Optimal transport at finite temperature

Patrice Koehl,[1] Marc Delarue,[2] and Henri Orland[3]
[1]*Department of Computer Science and Genome Center, University of California, Davis, California 95616, USA*
[2]*Unité de Dynamique Structurale des Macromolécules, Department of Structural Biology and Chemistry, UMR 3528 du CNRS, Institut Pasteur, 75015 Paris, France*
[3]*Institut de Physique Théorique, CEA-Saclay, 91191 Gif/Yvette Cedex, France*

(Received 29 April 2019; published 26 July 2019)

Optimal transport (OT) has become a discipline by itself that offers solutions to a wide range of theoretical problems in probability and mathematics. Despite its appealing theoretical properties, solving the OT problem involves the resolution of a linear program whose computational cost can quickly become prohibitive whenever the size of the problem exceeds a few hundred points. The recent introduction of entropy regularization, however, has led to the development of fast algorithms for solving an approximate OT problem. The successes of those algorithms have resulted in a popularization of the applications of OT in several applied fields such as imaging sciences and machine learning, and in data sciences in general. Problems remain, however, as to the numerical convergence of those regularized approximations towards the actual OT solution. In addition, the physical meaning of this regularization is unclear. In this paper, we propose an approach to solving the discrete OT problem using techniques adapted from statistical physics. Our first contribution is to fully describe this formalism, including all the proofs of its main claims. In particular we derive a strongly concave effective free energy function that captures the constraints of the optimal transport problem at a finite temperature. Its maximum defines a pseudo distance between the two set of weighted points that are compared, which satisfies the triangular inequalities. The temperature dependent OT pseudo distance decreases monotonically to the standard OT distance, providing a robust framework for temperature annealing. Our second contribution is to show that the implementation of this formalism has the same properties as the regularized OT algorithms in time complexity, making it a competitive approach to solving the OT problem. We illustrate applications of the framework to the problem of protein fold recognition based on sequence information only.

DOI: 10.1103/PhysRevE.100.013310

I. INTRODUCTION

Computing the distance between two probability distributions defined on a metric space M is a common problem in statistics. There are no single definitions of such a distance. Many statistical distances have been proposed, such as the total variation distance, the different divergence (Kullback-Leibler, Jensen-Shannon, etc.), distances based on energy, and so on. It is worth noting that many of those distances are not metrics. In addition, most of them only compute a single number when comparing two distributions. There are, however, many applications in which it is desirable to also generate a map, or "transport," between the two distributions of interest. If a cost is assigned to each of these possible maps, attempts to find the optimal map, namely the one with the lowest total cost, a problem referred to as the optimal transport (OT) problem, has enabled statisticians and mathematicians to derive a geometric structure on the space of probability distributions. The importance of this problem in those two fields may be best seen from the fact that two of its current main contributors have recently received Fields medals, Villani in 2010 and Figalli in 2018, in addition to Kantorovich receiving the Nobel prize in economics in 1975 for his contribution to optimal transport and its applications in economics. In addition, getting access to both the distance and the optimal transport map when comparing probability measures is of relevance to most, if not all, data science disciplines, and as such applications of OT have exploded in recent years, in domains such as machine learning [1], computer vision and image analysis [2–6], linguistics [7,8], differential geometry [9,10], geometric shape matching [11,12], and even music transcription [13], gene expression analyses [14], and the analysis of conformational dynamics of biomolecules [15]. Note that this is a small subset of all current applications, listed for illustration purpose only; for more extensive reviews of OT, we recommend [9,16–18] for reviews on the theory, [19,20] for reviews on its computational aspects, and [1] for a (brief) review of some of its applications.

The OT problem has been expressed in multiple forms, starting from the work of Monge in the 1780s [21], to be rediscovered or at least rephrased many times in the 1900s. For the sake of clarity, let us caricature it as follows: imagine we have N flour milling plants surrounding Paris, producing a total of 1 tonne of flour daily, and a distribution of P bakeries within Paris that consume a total of 1 tonne of flour each day. Knowing the cost $C(x, y)$ per unit weight of flour transported from a milling plant at x to a bakery at y, the problem is to define which milling plants should be supplying which bakeries so as to minimize the total transportation cost. In a more mathematical format, the milling plants and the bakeries lie in a metric space M. The flour production of the milling plants is represented by a probability measure μ, while

the flour consumption is represented by another probability measure ν. Let $C(x, y)$ be the cost of transporting flour from x to y, and $G(x, y)$ the amount of flour transported from x to y. G defines the transport plan. The optimal transport plan minimizes the total transportation cost U defined as

$$U(G) = \iint G(x, y)C(x, y)dxdy. \tag{1}$$

The minimum of $U(G)$ is to be found over the transport plans that satisfy the following constraints:

$$\forall x, y, \quad G(x, y) \geqslant 0, \tag{2a}$$

$$\forall x, \quad \int G(x, y)dy = \mu(x), \tag{2b}$$

$$\forall y, \quad \int G(x, y)dx = \nu(y). \tag{2c}$$

Constraint (2b) enforces that the total amount of flour delivered by plant x corresponds to its actual production, while constraint (2c) enforces that the total amount of flour delivered to bakery y corresponds to its actual need. The positivity constraint (2a) makes the problem physical. Finding a solution to the OT problem amounts to finding the optimal transport plan G_{opt}. The corresponding minimum transport cost U_{min} defines a "distance" between the two distribution measures μ and ν. The distance has all the properties of a metric when the cost matrix C is a metric matrix; see [17]. When $C(x, y) = d(x, y)^p$ where d is a metric of the space M, the distance is often referred to as the $p-$Wasserstein distance $W_p(\mu, \nu) = (U_{\text{min}})^{1/p}$ between the two measures. We note that when $p = 2$ and the cost matrix is based on the L_2 norm [i.e., $C(x, y)^p = ||x - y||^2$], the OT problem maps to the Schrödinger bridge problem [22], for which some simplifications are possible (see for example [5]). In this paper we will focus instead on the $1-$Wasserstein distance (i.e., with $p = 1$), also called the earth mover's distance, for a more general framework. Optimizing (1) under the constraints (2) is a linear programming (LP) problem. While much progress has been achieved for solving those problems [23], current practical implementations of algorithmic solutions are roughly of order $O(n^3)$, where n is the size of the discrete sets representing μ and ν, with a quadratic complexity in the number of variables considered. Such complexity levels are usually considered problematic when n is larger than a few thousands.

The current successes of OT did not come from recent improvements in solving LP problems. Instead, they have been triggered by the idea of minimizing a regularized version of Eq. (1):

$$\begin{aligned} U(\epsilon, G) &= U(G) - \epsilon H(G) \\ &= \iint G(x, y)C(x, y)dxdy \\ &\quad + \epsilon \iint G(x, y)\ln[G(x, y)]dxdy, \end{aligned} \tag{3}$$

where ϵ is the regularization parameter, and the second term $H(G)$ is an entropic barrier that enforces the positivity of the transport plan [24] (note that other penalty functions have been considered; see [20] for discussions). This regularized version of optimal transport is often called the Schrödinger problem [22]. It maps to the traditional OT problem as $\epsilon \to 0$; in addition, the optimal solution at a given ϵ defines a distance with metric properties, referred to as the Sinkhorn distance [24]. The entropic penalization has the advantage that it defines a strongly convex problem (as opposed to the original OT problem) with a unique solution [24]. Another advantage of the regularized OT problem is that its solution can be found efficiently through the so-called iterative proportional fitting procedure [25], also known as the Sinkhorn algorithm [26] or Sinkhorn-Knopp algorithm [27]. Many variants of those algorithms have been developed for solving regularized OT problems; we refer to [28–30] for overviews on those methods. Those algorithms find solutions for a given value of the relaxation parameter ϵ. For small values of this parameter, numerical issues can arise and a stabilization of the algorithm is necessary [31]. Despite such stabilization, convergence of a stabilized Sinkhorn-Knopp algorithm can nevertheless be very slow when ϵ is small. Such small values are, however, desirable for finding good approximations to the solution of the original nonregularized OT problem. A popular heuristic solution to this problem is the so-called ϵ scaling, where one subsequently solves the regularized problem with gradually decreasing values for ϵ (see for example [32]). To our knowledge, no quantitative analyses of the convergence of such an ϵ-scaling method are available. In particular, it is unclear whether the Sinkhorn distance is monotonic with respect to ϵ.

Our focus in this paper is on providing an alternate framework for solving the OT problem, as derived from a statistical physics point of view, in which we fully exploit the formal analogy of the cost function in Eq. (3) to a free energy, with ϵ an analog of a temperature, T. It can be seen as a generalization of the so-called invisible hand algorithm, which used a similar framework for solving the assignment problem in which the transportation plan G is encoded as a binary matrix [33]. This paper serves as a theoretical companion paper to Ref. [34], where we introduce the framework and apply it to the problem of defining a distance between 2D images. It provides the proofs of all the properties associated with the free energy we introduce, in particular its metric properties and its monotonic convergence to the "true" OT distance.

The paper is organized as follows. We start with a brief review of both the OT problem and its regularized version in the discrete case, with some proofs related to their metric properties. In Sec. III, we describe in detail the framework we propose for an optimal transport at finite temperature. Proofs of its major properties are provided in the appendices. The following section briefly describes the implementation of the framework in a C++ program, FreeOT. In Sec. V, we present applications of this framework to the problem of protein fold recognition based on sequence information only. We finally conclude with a detailed comparison between the entropy-regularized formulation of the OT problem and our formalism, as well as with a discussion on future developments.

II. THE REGULARIZED OPTIMAL TRANSPORT PROBLEM

This section provides a brief overview of the discrete optimal transport problem and its regularized version, covering

definitions as well as some considerations on its implementations. More thorough presentations can be found in Ref. [20].

We consider here the discrete version of the OT problem, i.e., optimal transport between two discrete probability measures. We consider two sets of points S_1 and S_2 of size N (for simplicity, we will assume that the two sets have the same size; note that the formalism can easily be extended to different sizes). Each point k in S_1 (resp. S_2) is assigned a "mass" $m_1(k)$ [resp. $m_2(k)$]. The balance condition implies that $\sum_k m_1(k) = \sum_l m_2(l)$. For simplicity, we assume that these two sums are equal to 1. We encode the cost of transport between S_1 and S_2 as a positive cost matrix C_{kl} with $(k,l) \in [1,N]^2$. The discrete optimal transport problem can then be formulated as finding a transport plan G, namely a matrix of correspondence between points k in S_1 and points l in S_2 that minimizes the total transport cost U defined as

$$U(G) = \sum_{k,l} G(k,l)C(k,l), \tag{4}$$

where the summations extend over all $(k,l) \in [1,N]^2$. The minimum of U is to be found for the matrices G that satisfy the following constraints:

$$\forall (k,l), \quad G(k,l) \geqslant 0, \tag{5a}$$

$$\forall k, \quad \sum_l G(k,l) = m_1(k), \tag{5b}$$

$$\forall l, \quad \sum_k G(k,l) = m_2(l). \tag{5c}$$

Note that the first condition, (5a), extends to $0 \leqslant G_{kl} \leqslant 1$ for all k and l, based on our assumption that the sum of the discrete probability measures are 1 on both sets of points. Matrices G that satisfy those conditions (5) belong to a polytope that we note as $\mathcal{G}(S_1, S_2)$.

The solution to this problem is an optimal transport plan G_{opt} and the corresponding minimum transport cost $d(S_1, S_2) = U(G_{\text{opt}})$. Note that this solution and its properties depend strongly on the choice of the cost matrix C. In particular, if we consider three sets of points S_1, S_2, and S_3, it is often of interest to have C satisfy metric properties, namely that

$$\begin{aligned} &\forall (k,j,l) \in [1,N]^3, \quad C(k,l) \leqslant C(k,j) + C(j,l), \\ &\forall (k,l) \in [1,N]^2, \quad C(k,l) = 0 \Leftrightarrow k = l. \end{aligned} \tag{6}$$

Villani [17] proved the following properties for $U_{\min}$:

Property 1. The optimal transport cost $d(S_1, S_2)$ is a distance between S_1 and S_2 that satisfies all axioms of a distance when C is a metric matrix, as defined above.

The gluing lemma [17] is the key to proving this property. As it will be used in the following, we write its discrete version here.

Lemma 1 (gluing lemma). Let S_1, S_2, and S_3 be three sets of points, with associated mass vectors $\mathbf{m}_1$, $\mathbf{m}_2$, and $\mathbf{m}_3$. Let $G_{12} \in \mathcal{G}(S_1, S_2)$ and $G_{23} \in \mathcal{G}(S_2, S_3)$ be two transport plans between S_1 and S_2, and between S_2 and S_3, respectively. Let G_{13} be the matrix defined by $G_{13}(k,l) = \sum_j \frac{G_{12}(k,j)G_{23}(j,l)}{m_2(j)}$. Then $G_{13} \in \mathcal{G}(S_1, S_3)$; i.e., G_{13} is a transport plan between S_1 and S_3.

Solving for the transport plan that minimizes Eq. (4) under the constraints (5) is a linear programming problem with an $O(N^3)$ complexity. To circumvent this large computing cost when N is large, Cuturi proposed to minimize a regularized version of Eq. (4):

$$U_\epsilon(G) = \sum_{kl} G(k,l)C(k,l) + \epsilon \sum_{k,l} G(k,l)\ln[G(k,l)], \tag{7}$$

where ϵ is the regularization parameter, and the second term is an entropic barrier that enforces the positivity of the G_{kl} terms [24]. With the addition of the entropic term controlling condition (5a), the two other conditions (5a) and (5b) are then enforced by introducing new auxiliary variables λ_k and μ_l as Lagrange multipliers,

$$\begin{aligned} \mathcal{L}_\epsilon = &\sum_{kl} G(k,l)C(k,l) - \epsilon \sum_{k,l} G(k,l)\ln[G(k,l)] \\ &- \sum_k \lambda_k \left[\sum_l G(k,l) - m_1(k) \right] \\ &- \sum_l \mu_l \left[\sum_k G(k,l) - m_2(l) \right]. \end{aligned} \tag{8}$$

Setting $\frac{\partial \mathcal{L}_\epsilon}{\partial G(k,l)} = \frac{\partial \mathcal{L}_\epsilon}{\partial \lambda_k} = \frac{\partial \mathcal{L}_\epsilon}{\partial \mu_l} = 0$, the critical points of the Lagrangian $\mathcal{L}_\epsilon$ satisfy the following conditions:

$$\begin{aligned} &G(k,l) = A_k \exp\left(-\frac{C(k,l)}{\epsilon}\right) B_l, \\ &\sum_l G(k,l) = m_1(k), \\ &\sum_k G(k,l) = m_2(l), \end{aligned} \tag{9}$$

where $A(k) = \exp(\frac{\lambda_k}{\epsilon} - 0.5)$ and $B(l) = \exp(\frac{\mu_l}{\epsilon} - 0.5)$. If we set K the matrix defined by $K_{kl} = \exp(-\frac{C_{kl}}{\epsilon})$, the conditions (9) can be rewritten in vector form as

$$\begin{aligned} &G = \text{diag}(\mathbf{A}) K \text{diag}(\mathbf{B}), \\ &\mathbf{A} \circ (K\mathbf{B}) = \mathbf{m}_1, \\ &\mathbf{B} \circ (K^T\mathbf{A}) = \mathbf{m}_2, \end{aligned} \tag{10}$$

where $\circ$ is the Hadamard (i.e., elementwise) product. Solving the OT problem by solving those equations leads to two main improvements compared to the standard linear optimization approach:

(i) The matrix G is directly computed from the vectors A and B; this leads to a reduction of the number of variables from N^2 to $2N$.

(ii) Equations (10) enable a simple iterative scheme to compute $\mathbf{A}$ and $\mathbf{B}$, namely, $(\mathbf{A}, \mathbf{B}) \leftarrow (\mathbf{m}_1 \oslash /(K\mathbf{B}), \mathbf{m}_2 \oslash /(K^T\mathbf{A}))$, where $\oslash$ is the Hadamard (elementwise) division. This iterative scheme is known as the Sinkhorn algorithm [26,27].

As discussed in the introduction, while these remarks lead to a significant reduction in computing time, there remain difficulties when solving the regularized OT problem when $\epsilon \to 0$, which is required to reach the true OT distance. This will be discussed further in the next section.

In contrast to the nonregularized transport cost U, the regularized transport cost U_ϵ [Eq. (4)] does not directly define a distance. However, it is possible to derive a distance from the regularized OT, using the following property (adapted from [24], in which a proof is provided):

Property 2. For $\epsilon > 0$, let G_ϵ^{opt} be the transport plan that minimizes the regularized transport cost $U_\epsilon(G)$ over all $G \in \mathcal{G}(S_1, S_2)$. Then $d_\epsilon(S_1, S_2) = \sum_{k,l} G_\epsilon^{\text{opt}}(k, l)C(k, l)$ is a distance between S_1 and S_2 that is symmetric and satisfies all triangular inequalities.

We conclude this section with a discussion of the relevance of Properties 1 and 2, namely that the solutions of the OT and regularized OT problems define a distance between the two discrete measures considered. The properties associated with distances are desirable, and we will ensure that we have them for whichever notion of similarity we introduce. As elegantly discussed by Mémoli [35], the triangular inequality properties of a distance d imply that if one is interested in comparing two continuous distributions λ and μ, and if S_1 and S_2 are finite supports to sample λ and μ, then

$$|d(\lambda, \mu) - d(S_1, S_2)| \leqslant d(\lambda, S_1) + d(\mu, S_2). \tag{11}$$

In practice we always have to rely on finite samples. It is clear that the quality of the approximation of a distribution λ by such a finite support S_1 is described by $d(\lambda, S_1)$. Therefore Eq. (11) indicates that comparing the discrete samples gives a measure of similarity of the underlying continuous distributions that is as good as how those discrete samples describe those distributions.

III. THE OPTIMAL TRANSPORT PROBLEM AT FINITE TEMPERATURE

Let us consider a system in thermal equilibrium at a finite temperature T. This system will sample several states, with each state characterized by a probability that is related to the energy of that state. The most probable state is the one with lowest energy. Using this framework from statistical physics, minimizing an energy function can be reformulated as the problem of finding the most probable state of the system it defines. Let us apply this framework to the discrete OT problem between two sets of points S_1 and S_2, using all the definitions from above. The "system" is then identified with the different transport plans between S_1 and S_2 equipped with masses $\mathbf{m}_1$ and $\mathbf{m}_2$, respectively, that satisfy the constraints of mass balance and positivity, namely that belong to $\mathcal{G}(S_1, S_2)$. We will slightly adapt the definition of this set by replacing the condition that a matrix G of this set must satisfy $0 \leqslant G_{kl}$ for all $(k, l) \in [1, N]^2$ with the condition that $0 \leqslant G_{kl} \leqslant 1$. The upper bound of 1 is a direct consequence of the fact that we impose $\sum_k m_1(k) = \sum_l m_2(l) = 1$.

Each state of the OT system is identified with a transport plan G, and its energy $U(G)$ is defined in Eq. (4). The probability distribution function for this system, $P(G)$, also referred to as the Gibbs distribution, is defined as

$$P(G) = \frac{1}{Z_\beta(S_1, S_2)} e^{-\beta U(G)}. \tag{12}$$

In this equation, $\beta = 1/(k_B T)$, where k_B is the Boltzmann constant and T the temperature, and $Z_\beta(S_1, S_2)$ is the partition function computed over all states of the system. This partition function is given by

$$Z_\beta(S_1, S_2) = \int_{G \in \mathcal{G}(S_1, S_2)} e^{-\beta U(G)} d\mu_{12}, \tag{13}$$

where $d\mu_{12}$ can be seen as the Lebesgue measure for the space of transport plans $\mathcal{G}(S_1, S_2)$. The partition function Z is related to the free energy of the system by

$$\mathcal{F}_\beta(S_1, S_2) = -\frac{1}{\beta} \ln[Z_\beta(S_1, S_2)] \tag{14}$$

and to the average energy $E_\beta(S_1, S_2) = \langle U(G) \rangle_{G \in \mathcal{G}}$ by

$$E_\beta(S_1, S_2) = -\frac{\partial \ln[Z_\beta(S_1, S_2)]}{\partial \beta}. \tag{15}$$

We note first two important properties of the free energy and average energy:

Property 3. For all $\beta > 0$, the free energy $\mathcal{F}_\beta(S_1, S_2)$ is symmetric and satisfies all triangle inequalities if the cost matrix C between S_1 and S_2 is metric.

Property 4. For all $\beta > 0$, the free energy $\mathcal{F}_\beta(S_1, S_2)$ and the average energy $E_\beta(S_1, S_2)$ are monotonically decreasing functions of β. Both converge to the traditional optimal transport distance $d(S_1, S_2)$.

Proof. The symmetry of $\mathcal{F}_\beta(S_1, S_2)$ is a direct consequence of the symmetry of the metric matrix C. The proof that it also satisfies all triangle inequalities is given in Appendix A, while the behavior of $\mathcal{F}_\beta$ and of E_β is analyzed in Appendix B. ■

This statistical physics formulation of the optimal transport problem is appealing. It defines a temperature dependent free energy that satisfies metric properties when the cost function is metric, with a monotonic dependence on the temperature (or inverse of the temperature, β), and convergence to the actual optimal transport distance at zero temperature. It is, however, of limited interest in practice as the partition function and therefore the free energy cannot be computed explicitly. We propose a scheme for approximating these quantities using the saddle point approximation. We will show that the corresponding mean field values satisfy properties similar to the exact quantities defined above. These mean field values can be readily computed.

Taking into account the constraints that define $\mathcal{G}(S_1, S_2)$, the partition function can be rewritten as

$$\begin{aligned} Z_\beta(S_1, S_2) = \int_0^1 \prod_{kl} dG(k, l) e^{-\beta \sum_{kl} C(k,l)G(k,l)} \\ \times \prod_k \delta\left(\sum_l G(k, l) - m_1(k)\right) \\ \times \prod_l \delta\left(\sum_k G(k, l) - m_2(l)\right). \end{aligned} \tag{16}$$

Using Fourier, we can represent a delta function as an integral of an exponential,

$$\delta(x) = \frac{1}{2\pi} \int e^{-ixt} dt, \tag{17}$$

where the integration is usually performed along the real axis. Introducing new auxiliary variables $\lambda(k)$ and $\mu(l)$, with

$(k,l) \in [1,N]^2$, and omitting the unessential normalization factors $1/(2\pi)$, the partition function can be written as

$$Z_\beta(S_1,S_2) = \int_0^1 \prod_{k,l} dG(k,l) e^{-\beta \sum_{kl} C(k,l)G(k,l)} \times \int \prod_k d\lambda(k) e^{-i\beta \sum_{k,l} \lambda(k)G(k,l) + i\beta \sum_k \lambda(k) m_1(k)} \times \int \prod_l d\mu(l) e^{-i\beta \sum_{k,l} \mu(l)G(k,l) + i\beta \sum_l \mu(l) m_2(l)}. \tag{18}$$

We have factored out β for the variables $\lambda(k)$ and $\mu(l)$ for consistency with the first term. Note that the integrand in Z is now a complex function, while Z itself is a real number. The imaginary part can be absorbed into λ and μ, i.e., $\lambda(k) \equiv i\lambda(k)$ and $\mu(l) \equiv i\mu(l)$, with now λ and μ being complex variables.

Rearranging the order of integration and reorganizing the exponential terms, we get

$$Z_\beta(S_1,S_2) = \int \prod_k d\lambda(k) \int \prod_l d\mu(l) \int_0^1 \prod_{k,l} dG(k,l) \, e^{-\beta \sum_{k,l} G(k,l)[C(k,l)+\lambda(k)+\mu(l)] + \beta\left(\sum_k \lambda(k) m_1(k) + \sum_l \mu(l) m_2(l)\right)}. \tag{19}$$

Performing the integration over the real variables $G(k,l)$ (most inner integrals), we get

$$Z_\beta(S_1,S_2) = \int \prod_k d\lambda(k) \int \prod_l d\mu(l) e^{\beta\left(\sum_k \lambda(k) m_1(k) + \sum_l \mu_l m_2(l)\right)} \times \prod_{kl} \frac{1 - e^{-\beta(C(k,l)+\lambda(k)+\mu(l))}}{\beta(C(k,l)+\lambda(k)+\mu(l))}. \tag{20}$$

We rewrite this partition function as

$$Z_\beta(S_1,S_2) = \int \prod_k d\lambda(k) \int \prod_l d\mu(l) e^{-\beta F_\beta(\boldsymbol{\lambda},\boldsymbol{\mu})}, \tag{21}$$

where $F_\beta(\boldsymbol{\lambda},\boldsymbol{\mu})$ is a functional, or effective free energy defined by

$$F_\beta(\boldsymbol{\lambda},\boldsymbol{\mu}) = -\left[\sum_k \lambda(k) m_1(k) + \sum_l \mu_l m_2(l)\right] - \frac{1}{\beta} \sum_{kl} \ln\left[\frac{1 - e^{-\beta(C(k,l)+\lambda(k)+\mu(l))}}{\beta(C(k,l)+\lambda(k)+\mu(l))}\right]. \tag{22}$$

Let $\bar{G}(k,l)$ be the expected value of $G(k,l)$ with respect to the Gibbs distribution given in Eq. (12). It is straightforward from the definition of the energy $U(G)$ and of the Gibbs distribution that

$$\bar{G}(k,l) = -\frac{1}{\beta} \frac{\partial Z_\beta(S_1,S_2)}{\partial C(k,l)}. \tag{23}$$

It is unfortunately not possible to compute these expected values directly from this equation, as the partition function is not known analytically. Instead, we derive a saddle point approximation (SPA). The SPA is computed by looking for extrema of the effective free energy with respect to the variables $\lambda(k)$ and $\mu(l)$:

$$\frac{\partial F_\beta(\boldsymbol{\lambda},\boldsymbol{\mu})}{\partial \lambda(k)} = 0 \quad \text{and} \quad \frac{\partial F_\beta(\boldsymbol{\lambda},\boldsymbol{\mu})}{\partial \mu(l)} = 0. \tag{24}$$

After some rearrangements, those two equations lead to the following system of equations:

$$\bar{G}(k,l) = \phi(\beta(C(k,l) + \lambda(k) + \mu(l))), \tag{25}$$

$$\sum_l \bar{G}(k,l) = m_1(k), \tag{26}$$

$$\sum_k \bar{G}(k,l) = m_2(l), \tag{27}$$

where

$$\phi(x) = \frac{e^{-x}}{e^{-x} - 1} + \frac{1}{x}. \tag{28}$$

Note that $\phi(x)$ is related to the Langevin function $L(x)$ by $\phi(x) = \frac{1}{2}[1 - L(\frac{x}{2})]$. This function $\phi(x)$ is defined and continuous for all real values x [with the extension that $\phi(0) = 0.5$], monotonically decreasing over $\mathbb{R}$, with asymptotes $y = 1$ and $y = 0$ at $-\infty$ and $+\infty$, respectively [see Appendix E for a representation of $\phi(x)$]. As such, it correctly constrains the values of the transport plan G to be in the range of values [0,1].

One can see that the variables $\lambda(k)$ and $\mu(l)$ must be real as the transport plan is real. Another way to see this is to recognize that the complex integral defining the partition function [see Eq. (20)] does not depend on the choice of the integration paths. The saddle point equations (27) indicate that a path parallel to the real axis for each of the variables $\lambda(k)$ and $\mu(l)$ is preferred.

We observe that Eqs. (22) and (27) are invariant under the constant translation $\{\lambda(k) + K, \mu(l) - K\}$, where K is an arbitrary real constant. This translational degree of freedom leaves the effective free energy $F_\beta(\boldsymbol{\lambda},\boldsymbol{\mu})$ unchanged. To analyze the validity of the saddle point approximation, we need to check the existence and assess the unicity of the critical points of this effective free energy. The following theorem shows that $F_\beta(\boldsymbol{\lambda},\boldsymbol{\mu})$ is weakly concave and can be made strictly concave on a subspace of the parameter space that is easily defined.

Theorem 1. The Hessian of the effective free energy $F_\beta(\boldsymbol{\lambda},\boldsymbol{\mu})$ is negative semidefinite with $(2N-1)$ negative eigenvalues and one zero eigenvalue. Furthermore, the eigenvector corresponding to the zero eigenvalue is $(1,\ldots,1,-1,\ldots,-1)$ (with N 1's, and N -1's), and thus corresponds to the constant translation invariance of this energy. Setting one of the parameters $\lambda(k)$ or $\mu(l)$ as zero, the free energy function on this restricted parameter space is strictly concave.

Proof. See Appendix C. ∎

For a given value of the parameter β, the expected values $\bar{G}(k,l)$ that are solutions to the system of equations (27) form a transport plan G_β^{opt} between S_1 and S_2 that is optimal with respect to the free energy defined in (22). We can associate with this transport plan an optimal free energy F_β^{MF} and an optimum energy $U_\beta^{MF} = \sum_{k,l} G_\beta^{\text{opt}}(k,l) C(k,l)$. Note that those two values are the mean field approximations of the

exact free energy and internal energy defined in Eqs. (14) and (15), respectively. We now list important properties of U_β^{MF} and F_β^{MF}.

Property 5. For all $\beta > 0$ and cost metric matrix C, U_β^{MF} is symmetric and satisfies all triangle inequalities.

Proof. The symmetry of U_β^{MF} is a direct consequence of the symmetry of the metric matrix C. The proof for the triangle inequalities is given in Appendix D. ■

Property 6. F_β^{MF} and U_β^{MF} are monotonic decreasing functions of the parameter β. In addition, both converge to the optimal transport energy defined in Eq. (1).

Proof. See Appendix E. ■

Theorem 1 and the two Properties 5 and 6 highlight a number of advantages of the proposed framework that rephrases the optimal transport problem as a temperature dependent process. First, at each temperature the optimal transport problem is turned into a strongly concave problem with a unique solution. This problem has a linear complexity in the number of variables, compared to the quadratic complexity of the original problem. The concavity allows for the use of simple algorithms for finding a minimum of the effective free energy function [Eq. (22)]. We note also that Eqs. (27) provide good numerical stability for computing the transport plan, because of the ratio of exponentials. Second, the modified problem defines an optimal distance at each temperature, that converges to the traditional optimal transport distance when $T \to 0$. Finally, the convergence as a function of temperature is monotonic.

IV. IMPLEMENTATION

We have implemented the finite temperature optimal transport framework described here in a C++ program FreeOT that is succinctly described in Algorithm 1.

Algorithm 1. *FreeOT: a temperature dependent framework for computing the optimal transport distance between two weighted set of points.*

Input: The two sets of points S_1 and S_2, and their associated weights $\mathbf{m}_1$ and $\mathbf{m}_2$. Cost matrix C between S_1 and S_2. Initial value β_0 for β

Initialize: Initialize arrays $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$ to 0. Set $STEP = \sqrt{10}$. Set $\beta^0 = \beta_0/STEP$

for $k = 1, \ldots$ until convergence **do**
(1) Initialize $\beta^k = STEP * \beta^{k-1}$
(2) Solve nonlinear equations (27) at saddle point
(3) Compute optimal transport plan G_β^{opt} and $U^{MF}(\beta^k)$
(4) Check for convergence: if $|U^{MF}(\beta^k) - U^{MF}(\beta^{k-1})|/U^{MF}(\beta^{k-1}) < TOL$, stop
end for

Output: The converged transport plans $G_\beta^{\text{opt}}(k, l)$ and the corresponding transport costs $U^{MF}(\beta)$.

FreeOT is based on an iterative procedure in which the parameter β (inverse of the temperature) is gradually increased. At each value of β, the nonlinear system of equations defined by Eq. (27) is solved using an iterative Newton-Ralphson method. At each iteration for this Newton method, the Jacobian of the system of equations is computed, and a linear system is solved based on this Jacobian, whose solution provides estimates for the arrays of parameters $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$. These new estimates are then used to assess how well the SPA equations are satisfied. Once the errors on the SPA equations fall below a tolerance TOL (usually set to 10^{-8}), the optimal transport plan G_β^{opt} and the corresponding transport energy $U^{MF}(\beta)$ are computed. If the latter falls within the tolerance TOL of the corresponding value computed for the previous β value, the procedure is deemed to have converged and the program is stopped. Note that the converged values of $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$ at a given β serve as input for the following β.

The time complexity of FreeOT is dominated by step (2) in its algorithm, namely solving the nonlinear system equations defined by the SPA. Let us rewrite the saddle point equations (27) as functions of the parameters $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$ only:

$$\begin{aligned}\sum_l \bar{G}(k,l) &= \sum_l \frac{\exp -\beta(C(k,l)+\lambda(k)+\mu(l))}{\exp -\beta(C(k,l)+\lambda(k)+\mu(l))-1} \\ &\quad + \sum_l \frac{1}{\beta(C(k,l)+\lambda(k)+\mu(l))} = m_1(k), \\ \sum_k \bar{G}(k,l) &= \sum_k \frac{\exp -\beta(C(k,l)+\lambda(k)+\mu(l))}{\exp -\beta(C(k,l)+\lambda(k)+\mu(l))-1} \\ &\quad + \sum_k \frac{1}{\beta(C(k,l)+\lambda(k)+\mu(l))} = m_2(l).\end{aligned} \tag{29}$$

Let us then define

$$\begin{aligned}A_\lambda(k) = &- \sum_l \frac{\exp -\beta(C(k,l)+\lambda(k)+\mu(l))}{\exp -\beta(C(k,l)+\lambda(k)+\mu(l))-1} \\ &+ \sum_l \frac{1}{\beta(C(k,l)+\lambda(k)+\mu(l))} + m_1(k)\end{aligned} \tag{30}$$

and

$$\begin{aligned}A_\mu(l) = &- \sum_k \frac{\exp -\beta(C(k,l)+\lambda(k)+\mu(l))}{\exp -\beta(C(k,l)+\lambda(k)+\mu(l))-1} \\ &+ \sum_k \frac{1}{\beta(C(k,l)+\lambda(k)+\mu(l))} + m_2(l).\end{aligned} \tag{31}$$

The SPA equations become

$$\begin{aligned}A_\lambda(k) &= 0, \quad \forall k, \\ A_\mu(l) &= 0, \quad \forall l.\end{aligned} \tag{32}$$

Those equations form a system of $2N-1$ equations with $2N-1$ variables, N $\lambda(k)$ values, and $N-1$ $\mu(l)$ values [as a reminder $\mu(N)$ is set to zero to ensure that the free energy functional is concave]. Let us assume that we know an initial solution $\mathbf{X}_0 = (\boldsymbol{\lambda}_0, \boldsymbol{\mu}_0)$ for this system. Taylor expansions of the predicates $(\mathbf{A})$ in the neighborhood of this solution lead to the following system of equations:

$$J(\mathbf{X}_0)\delta\mathbf{X} = -\mathbf{A}(\mathbf{X}_0), \tag{33}$$

where $\delta\mathbf{X} = (\delta\boldsymbol{\lambda}, \delta\boldsymbol{\mu})$ is the correction to be applied to $\mathbf{X}_0$, $\mathbf{A}(\mathbf{X}_0)$ is the vector of values of the $2N-1$ predicates $(\mathbf{A}_\lambda, \mathbf{A}_\mu)$ at $\mathbf{X}_0$, and $J(\mathbf{X}_0)$ is the Jacobian of $\mathbf{A}$ taken at $\mathbf{X}_0$. We note that this Jacobian J is equal to the opposite of the Hessian

of the free energy function F. As this free energy is concave, the Jacobian is then positive definite. It can be written in block form:

$$J(\mathbf{X}_0) = \begin{bmatrix} D_\lambda & G' \\ G'^T & D_\mu \end{bmatrix}, \tag{34}$$

where $G'(k,l) = \beta\phi'(\beta(C(k,l) + \lambda(k) + \mu(l)))$, D_λ is the diagonal matrix defined by $D_\lambda(k,k) = \sum_l G'(k,l)$, D_μ is the diagonal matrix defined by $D_\mu(l,l) = \sum_k G'(k,l)$, and $\phi'(x)$ is the derivative of the function $\phi(x)$ defined in Eq. (28) (see Appendix C). The system of equations (33) can then be rewritten as

$$\begin{bmatrix} D_\lambda & G' \\ G'^T & D_\mu \end{bmatrix} \begin{bmatrix} \delta\boldsymbol{\lambda} \\ \delta\boldsymbol{\mu} \end{bmatrix} = -\begin{bmatrix} \mathbf{A}_\lambda \\ \mathbf{A}_\mu \end{bmatrix} \tag{35}$$

or equivalently as:

$$\begin{cases} D_\lambda \delta\boldsymbol{\lambda} + G'\delta\boldsymbol{\mu} &= -\mathbf{A}_\lambda, \\ G'^T\delta\boldsymbol{\lambda} + D_\mu\delta\boldsymbol{\mu} &= -\mathbf{A}_\mu. \end{cases} \tag{36}$$

Multiplying the bottom equation by $G'D_\mu^{-1}$ and subtracting from the top equation, we get

$$\left(D_\lambda - G'D_\mu^{-1}G'^T\right)\delta\boldsymbol{\lambda} = \mathbf{A}_k - G'D_\mu^{-1}\mathbf{A}_l. \tag{37}$$

Once this system is solved for $\delta\boldsymbol{\lambda}$, we can solve for $\delta\boldsymbol{\mu}$ using the equation $G'^T\delta\boldsymbol{\lambda} + D_\mu\delta\boldsymbol{\mu} = \mathbf{A}_\mu$. Note that $D_\lambda - G'D_\mu^{-1}G'^T$ is the Shur complement of D_μ in J. Using this representation reduces the problem of solving a system of size $(2N-1)\times(2N-1)$ to that of solving a system of size $N \times N$. Note that since the Jacobian J is positive definite, the Shur complement of D_μ in J is also positive definite. To solve the system in Eq. (37), we have implemented both a direct method based on an LDL^T decomposition of the Shur complement and an iterative method based on conjugate gradient. The performances of these two methods will be compared below in computational experiments.

V. COMPUTATIONAL EXPERIMENTS

A. Comparing protein sequences using finite temperature optimal transport

We present some computational examples that illustrate the use of our framework. We consider the problem of comparing protein sequences. A protein sequence is usually represented as a string of letters, where each letter corresponds to an amino acid. This representation has proved very useful, especially in the context of sequence alignment [36,37] that is usually performed using string-matching algorithms [38]. When comparing two sequences, these algorithms proceed in two steps, first the generation of the alignment between the two sequences, then the derivation of a statistical score for that alignment. It should be noted that this score is not a metric in sequence space. "Alignment-free" methods have been proposed as an alternate solution to measure the similarity of two protein sequences that enforce the metric property (for a review, see [39–42]). Most of these methods compute the frequencies of words of a fixed length, k, also denoted as *k-mers*. Once the frequency distribution functions of such k-mers have been computed for two sequences, the distance between those two sequences is assigned to be the distance between those distributions [40,43]. The finite temperature optimal transport framework allows us to combine the benefits of those two approaches. It is adapted to comparing protein sequences as follows. We first consider a kernel for amino acid pairs, namely a symmetric, positive definite matrix K_1 such that $K_1(i,j)$ gives a quantitative value for the similarity between amino acid of type i and amino acid of type j. To build such a kernel, we consider the matrices representing the raw data *SM* of any BLOSUM matrices, namely the raw count of how often an amino acid of type i is substituted by amino acid j in a set of selected protein sequence alignments [44]. This matrix is normalized by considering its row sums $P(i)$:

$$P(i) = \sum_{j=1}^{20} SM(i,j),$$

$$SM2(i,j) = \frac{SM(i,j)}{P(i)p(j)}. \tag{38}$$

We have checked that when *SM* is a raw count BLOSUM matrix, then *SM*2 is symmetric, positive, and definite. Smale and colleagues [45] noticed that for a strictly positive real number β, the matrix K_1 defined as

$$K_1(i,j) = SM2(i,j)^\beta \tag{39}$$

is also symmetric, positive, and definite. In the following, we will use the BLOSUM62 matrix, with each element raised to the power 0.1, as suggested by Smale *et al.* [45].

The second step is to define a kernel for comparing two k-mers, namely two strings of the same length, k. Let $S_k = (s_1,\ldots,s_k)$ and $T_k = (t_1,\ldots,t_k)$ be such strings. The function K_2 defined by

$$K_2(S_k,T_k) = \prod_{l=1}^{k} K_1(s_l,t_l) \tag{40}$$

is a kernel on the space of strings of length k. This kernel is normalized,

$$\hat{K}_2(S_k,T_k) = \frac{K_2(S_k,T_k)}{\sqrt{K_2(S_k,S_k)K_2(T_k,T_k)}}, \tag{41}$$

and converted into a distance, or cost C, between S_k and T_k using

$$C(S_k,T_k) = \sqrt{2 - 2\hat{K}_2(S_k,T_k)}. \tag{42}$$

A pair of sequences S_1 and S_2 is represented with their sets of k-mers, the cost matrix C between those k-mers, with C computed as described above. The masses of the k-mers are set uniform. The k-mers are contiguous stretches of sequences; i.e., we do not consider gaps. In addition, k-mers may be overlapping; i.e., there are $N-k+1$ k-mers of length k for a sequence with length N.

We focus on classifying proteins into structural folds based on sequence information only. We considered protein sequences from the SCOPe/ASTRAL database [46]. The SCOPe database is designed to provide a comprehensive description of the structural and evolutionary relationships between all proteins whose structure is known. It defines a classification of those protein structures at four levels, namely class, folds, superfamilies, and families. Here we only

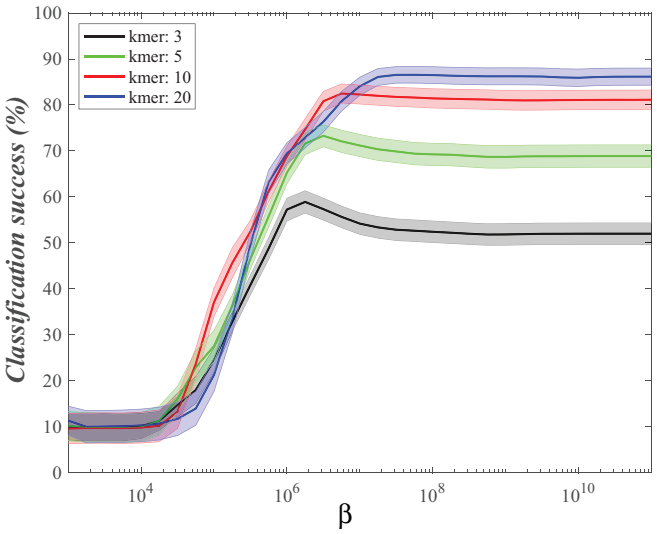

FIG. 1. Discriminative power of the temperature-based optimal transport distances for protein fold recognition. The probability of correct classification of a protein sequence into its fold defined by SCOPe based on the distance measure $D(\beta) = U_\beta^{MF}$ (see text for details) is plotted against $\beta = 1/T$ for different sizes of the k-mers used to represent the sequences. All the curves are arithmetic means over 10 000 classification experiments (see text for details). Shaded areas represent standard deviations.

TABLE I. Classification powers of different distances between protein sequences.

Distance	SCOPe Class P (SD)[a]	SCOPe Fold P (SD)[a]
OT k-mer 1[b]	50.0 (2.4)	31.2 (2.2)
OT k-mer 5	77.2 (2.2)	68.9 (2.4)
OT k-mer 10	86.7 (1.9)	81.1 (2.1)
OT k-mer 20	90.3 (1.7)	86.2 (1.8)
OT k-mer 30	90.0 (1.8)	86.0 (2.0)
FASTA[c]	91.0 (1.6)	89.1 (1.6)
Bray-Curtis k-mer 1[d]	48.0 (2.1)	29.0 (2.3)
Bray-Curtis k-mer 5	57.2 (2.4)	45.6 (2.3)
Jaccard k-mer 1	48.0 (2.4)	30.0 (2.2)
Jaccard k-mer 5	55.0 (2.4)	43.0 (2.3)

[a]Mean and standard deviation SD (in %) of the probability of correct classification at the level considered, computed over 10 000 classification experiments.
[b]Converged OT distance.
[c]The FASTA "distance" between two sequences is set to the raw score of the alignment of the two sequences, using BLOSUM62 as a substitution matrix, and gap penalties of −11 for opening and −1 for extension.
[d]Alignment-free "distances" between two sequences computed as the dissimilarities between the frequencies of their k-mer types. These distances were computed using the program Alfree [42].

considered the first two levels, as they are directly related to structures. ASTRAL is a compendium to SCOPe that provides databases of protein sequences and/or structures, as well as tools useful for analyzing protein structures and their sequences. We used a representative subset of the current SCOPe/ASTRAL database that contains protein sequences sharing less than 40% identity. This subset includes 12 different folds, three for each of the four classes. Using the SCOP terminology, we considered 3 mainly helical folds, a.1, a.25, a.121; 3 mainly β folds, b.6, b.29, b.42; 3 α/β folds, c.1, c.66, c.69; and three $\alpha + \beta$ folds, d.17, d.38, d.108; with 48, 46, 34, 43, 51, 30, 60, 41, 63, 46, 60, and 80 representatives, respectively, for a total of 602 protein sequences. We refer to this set as SCOP12.

We computed a set of matrices $D(\beta)$ for β ranging between 1000 and 10^{10}, such that $D(\beta)(k, l)$ is the optimized transport energy U_β^{MF} between the two sequences S_k and S_l in the set SCOP12. We also computed $D(\infty)$, namely the matrix of distances at convergence (usually reached for $\beta > 10^9$).

In order to assess the discriminative power contained in the different distance matrices $D(\beta)$, we considered a set of classification tasks as follows: We randomly selected half the sequences from each fold to form a training set and use it for performing first-nearest-neighbor classification [where nearest is with respect to the distance $D(\beta)$] to the remaining sequences. By simple comparison between the class predicted by the classifier and the actual class to which the image belongs we obtain an estimate of the probability of correct classification $P(\beta)$ using $D(\beta)$. We then repeat this procedure for 10 000 random choices of the training set. In Fig. 1, we plot $P(\beta)$ as a function of β for classification at the fold level for different values of the size of the k-mers. Note that the lower the temperature (or alternatively the higher the parameter β), the more discriminative the distance U_β^{MF}. The highest level of correct classification is already obtained for $\beta = 10^7$ for all values of k-mers, i.e., much before convergence to the optimal transport distance, usually reached for $\beta > 10^9$. In addition, the discriminative power of the temperature-based OT distance improves as the size of the k-mers representing the sequences increases.

In Table I, we report the probabilities of correct classifications for $D(\infty)$ at the SCOPe class and fold levels at different k-mer sizes, and compare them with the success rates of the alignment-based method FASTA [47] and of two alignment-free methods that compare the distributions of k-mers using either Jaccard index distance or the Bray-Curtis dissimilarity [48].

FASTA is a standard procedure in bioinformatics for comparing protein sequences that is based on dynamic programming. It proceeds in two steps, first with the generation of the alignment between the two sequences, then with the derivation of a score for that alignment. It relies on a weighting scheme to measure the cost of matching pairs of amino acids. Many such weights have been proposed, from substitution matrices such as the BLOSUM matrices [44], to matrices that capture physicochemical properties of amino acids [49]. Using this score, an alignment is derived following a dynamic programming algorithm, either the local method of Smith and Waterman [50] or the global method of Needleman and Wunsch [51]. This alignment is then scored by summing the individual weights of the matching pairs of amino acids and adding penalties for the presence of gaps. In our experiments, we have used the BLOSUM62 matrix, for consistency with the results based on optimal transport (see above), and gap penalties of −11 for opening and −1 for extension (the default

values when using the BLOSUM62 matrix). We used the SSEARCH tool within FASTA that is based on the Smith and Waterman dynamic programming method. The "distance" between two sequences is then set to the raw score of the alignment. It should be noted that this score is not a metric in sequence space.

As alternates to dynamic programming methods such as FASTA, many "alignment-free" methods have been proposed over the past three decades (for a review, see [39–41]). Most of these methods compute first the frequencies of words of a fixed length within a protein sequence, k, usually denoted as k-mers. Once the frequency distribution functions of such k-mers have been computed for two sequences, the distance between those two sequences is assimilated to the distance between those distributions, using different definitions of distance [40]. We have considered two such methods based on two different distances, the Jaccard index distance and the Bray-Curtis dissimilarity. The Jaccard distance is based on the presence or absence of k-mer types in the sequences. Briefly, let us consider two sequences S_1 and S_2, and let us consider that there are N types of possible k-mers in each of these sequences (for example, if $k = 2$ then $N = 20^2$). We then compare the two sequences by computing four indices M_{11}, M_{10}, M_{01}, and M_{00}, representing the number of types of k-mers that are found in S_1 and S_2, in S_1 but not in S_2, in S_2 but not in S_1, and neither in S_1 nor in S_2, respectively. The Jaccard distance between the 2 sequences is then

$$d_J(S_1, S_2) = \frac{M_{01} + M_{10}}{M_{01} + M_{10} + M_{11}}.$$

In contrast to the Jaccard distance, the Bray-Curtis dissimilarity measure takes onto account the actual number of k-mers of each type. If there are X_i k-mers of type i in S_1 and Y_i k-mers of the same type i in S_2, then

$$d_{BC}(S_1, S_2) = 1 - 2\frac{\sum_i \min(X_i, Y_i)}{\sum_i (X_i + Y_i)}.$$

Note that the Jaccard distance induces a metric on the sequence space, while the Bray-Curtis distance does not. We used the program Alfree [42] to compute these string-based, alignment-free distances between sequences.

In Table I, we report the probabilities of correct classifications for $D(\infty)$ at the SCOPe class and fold levels at different k-mer sizes, and compare them with the success rates of the alignment-based method (FAST) and of two alignment-free methods (Jaccard and Bray-Curtis) presented above. As already illustrated in Fig. 1, the discriminative power of the OT distance increases as the size of the k-mers increases up to 20, and reaches a plateau after that. The corresponding optimal OT distance for k-mers of size 20 is equivalent to the discriminative power of the Smith and Waterman alignment method. The significant difference however is that the OT distance is an actual distance, while the Smith and Waterman score is not. The two alignment-free methods based on k-mer frequencies within the sequences show significantly lower performances on this data set.

The experiments described above highlight the classification powers of the finite temperature OT distance that we have introduced. Interestingly, when we compute the OT distance between two sequences, we also derive the optimal transport plan between the k-mers of those two sequences. In Fig. 2, we show visual representations of the optimal transport plan G for the optimal transport between the sequence of concanavalin A (PDB code 2cnaA, 237 amino acids) and of a peanut lectin (PDB code 2pelA, 232 amino acids) for two k-mer sizes, 3 and 10. The FASTA alignment between those two sequences does not identify a single alignment; instead it finds that region 1-115 of concanavalin A aligns well with region 114:229 of lectin A, while region 2-102 of lectin A aligns well with region 124-227 of concanavalin A: there has been a domain swap between the two sequences. Clearly, the transportation plan between the two sequences captures that domain swap, especially for large k-mer sizes. This visualization of the transport plan is akin to the concept of dot plot representation of the similarity between two sequences [52].

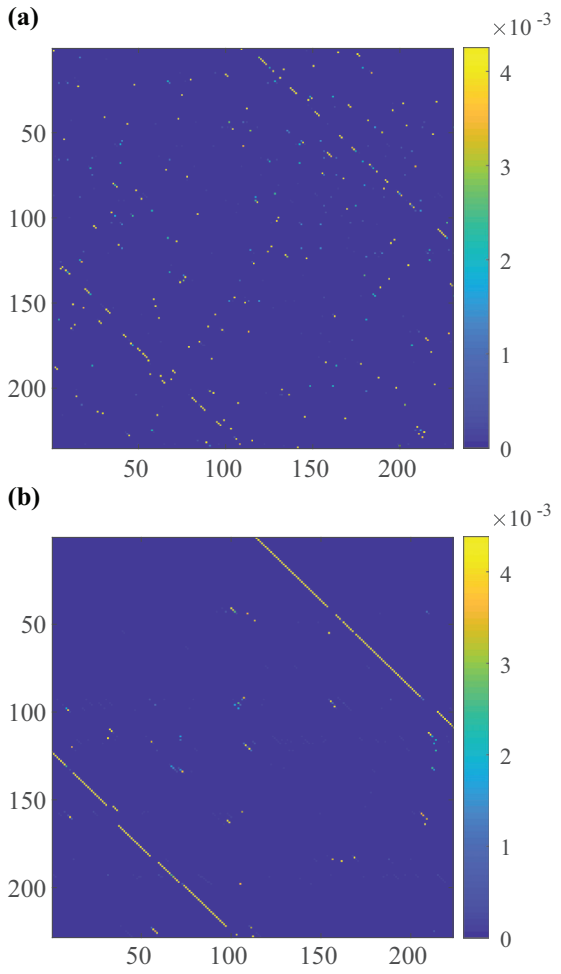

FIG. 2. Optimal transport plans between concanavalin A (horizontal axis) and peanut lectin (vertical axis), for k-mer sizes of 3 (a) and 10 (b). The local alignments of the two domains of these sequences, as well as the domain swap, appear clearly from the transportation plans.

B. Computing time

As described in the Implementation section above, the main computing cost of our implementation of the finite temperature optimal transport problem, FreeOT, is associated

with solving the nonlinear set of equations corresponding to the SPA at each value of β. We solve this system of equation using an iterative Newton-Ralphson method. At each iteration, we solve a linear system of equation based on the Jacobian of the nonlinear equations. As described in the Implementation section, this system can be rearranged to be of size $N \times N$, where N is the number of points considered. We considered two methods for solving this system. First, we use a direct method with which we decompose the matrix describing the system using an *LDL* decomposition, as implemented in the program "dsysv" from the LAPACK packages [53]. The corresponding time complexity is expected to be $O(N^3)$. Second, we implemented an iterative conjugate gradient (CG) method. Each iteration of the CG methods involves two matrix-vector multiplications, which are of order $O(N^2)$. The CG method will converge in at most N iterations, and in many cases in many less iterations. As such, it is expected to be faster than the direct method if the total number of CG iterations is small. We refer to these two implementations as FreeOT(direct) and FreeOT(iter).

Based on Theorem 1 and the two Properties 5 and 6, FreeOT is expected to provide a fast and robust solution to the OT problem. To check that this is indeed the case, we have compared FreeOT with our own implementation of the entropy-regularized approach to the OT problem. The latter, dubbed EntropyOT, is based on a log-domain stabilization and eta-scaling heuristic [32] and an overrelaxation scheme [54]. These two modifications to the original algorithm of Cuturi [24] are expected to improve convergence of the iterative scaling algorithm, as well as robustness for small values of the relaxation parameter ϵ through the use of logarithmic stabilization. We have experimented with applications of FreeOT and EntropyOT to compare protein sequences, as described above. We have compared each sequence in the SCOP12 data set defined above against five other sequences of similar lengths. The computing time for one sequence is then reported as the average over those five neighbors. Each comparison is made based on the BLOSUM62 matrix, with the size of the k-mers set to 1. The optimization is performed until convergence, i.e., until the relative change in the energy falls below a tolerance of 10^{-6}. Such convergence is usually reached for $\beta = 10^{11}$ (or equivalently for $\epsilon = 10^{-11}$ for EntropyOT). All computational experiments were performed on an iMac computer with a 4.0 GHz Intel Core I7 processor, with 64 GB of memory. The computing times for FreeOT (both the direct and iterative versions) and EntropyOT are plotted against the sizes of the protein sequences in Fig. 3.

With the exceptions of only small sequences, both versions of FreeOT are found to be faster than EntropyOT. We have assigned this difference to the fact that EntropyOT was found to slow down significantly for very small ϵ values. While convergence with high precision may not be needed, we observe that FreeOT is free of those convergence problems.

The running times for FreeOT based on the direct linear solver are consistent with an $O(N^3)$ time complexity, while the equivalent running times for the iterative CG solver are consistent with an $O(N^2)$. Note that those running times are reported for the full procedure that includes scaling the β parameter from a small value (high temperature) to a large value (low temperature). Interestingly, the differences

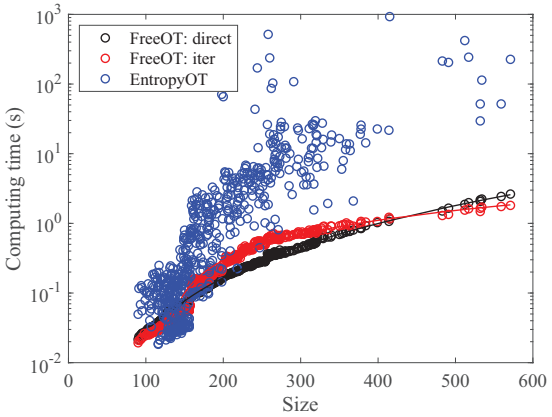

FIG. 3. Time complexity for FreeOT and EntropyOT. The running times for the finite temperature procedure (FreeOT) and for the entropy-regularized procedure (EntropyOT, blue circles) used for comparing two sequences are plotted against the size of the sequences, including two options for the inner solver of the linear system of equations (see text for details), namely a direct solver (black circles) or an iterative conjugate gradient solver (red circles). The corresponding solid lines shows the best fit to a cubic polynomial for the direct solver data, and to a quadratic polynomial for the iterative solver. The timings are computed on a single Intel Core I7 processor running at 4.0 GHz with 64 GB of RAM.

between the two solvers is small for the sizes of protein sequences considered here. In fact, up to size 400, the direct solver is found to be faster than the iterative solver. This is due to the fact that we are using an efficient, parallelized version of "dsysv" from LAPACK that uses all 8 cores available on the computer on which we ran those experiments. For those small values of N, the apparent time complexity of the parallelized direct solver is of order $O(N^2)$. When N becomes larger, the apparent complexity becomes closer to $O(N^3)$, and then the iterative solver becomes faster.

Both FreeOT and EntropyOT include a scaling of their regularization parameter, β and ϵ, respectively. This scaling is akin to an annealing procedure. As the values of the variables $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$ at one value of the regularization parameter are used as input to the next value of the regularization parameter considered, it is expected that convergence at this new step will be faster. To check whether this is true, we repeated those calculations by resetting the variables to 0 at each step, and compared the number of iterations needed to converge at each value of the relaxation between the scaling version, and reset version of FreeOT and EntropyOT. Most experiments using EntropyOT failed due to numerical instabilities for $\epsilon < 10^{-5}$. In contrast, FreeOT was able to converge even with reset of the variables, over the whole range of β values. The average numbers of iterations for the regular and reset version of FreeOT (both based on the iterative solver) are shown in Fig. 4.

For the regular version of FreeOT we see significant fluctuations for the number of iterations over the 602 sequences considered for β in the range $[10^4, 10^7]$. Above 10^7, this number remains small and constant (5). In comparison, the number of iterations needed at each β step increases as β increases when the variables are reset for each β value

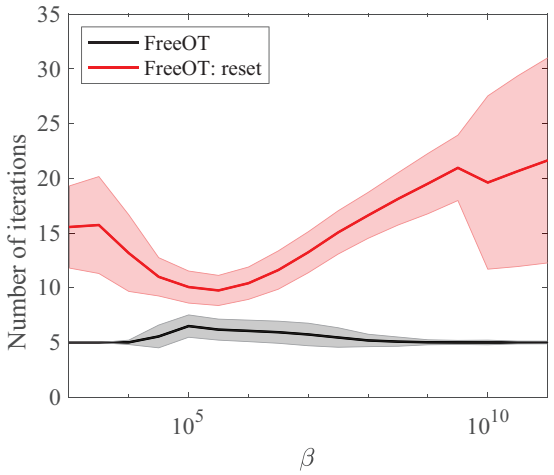

FIG. 4. Convergence of the nonlinear solver in FreeOT for successive values of β. The number of Newton-Ralphson iterations needed for solving the SPA conditions is plotted against the value of β, for the standard implementation of FreeOT with transfer of variables between two consecutive β values (in black), and for the reset version of FreeOT in which the variables are reset to 0 at each value of β. The solid line corresponds to the arithmetic means over the 602 sequence comparison experiments (see text for details). Shaded area represents standard deviations.

[FreeOT(reset)]. The ranges of values observed over all experiments are large for large values of β, highlighting difficulties to converge for those values. Notwithstanding, all computations converge, even with $\beta = 10^{11}$, thereby validating the stability of FreeOT.

VI. DISCUSSION

In this paper, we have proposed a statistical physics framework to solve the discrete optimal transport problem. Given two sets of weighted points S_1 and S_2, and a cost matrix between those sets, assumed to be metric, we have shown first that the free energy computed over the polytope of all possible transport plans between those two sets defines a temperature dependent distance between the sets that satisfies the symmetry and triangular inequality properties of a metric. While the free energy cannot be computed exactly, it can be estimated using a saddle point approximation. The saddle point approximation is derived by constructing a weakly concave effective energy function that captures the constraints of the optimal transport problem. This effective energy function is parameterized by temperature. Its maximum defines an optimal transport plan. We have shown that the transportation energy corresponding to this transport plan defines a temperature dependent distance between the two sets of points considered. We proved also that this energy decreases monotonically as a function of β (the inverse of temperature) to the standard optimal energy distance, providing a robust framework for temperature annealing. We described an application of our framework in bioinformatics, in which we have rephrased the problem of comparing two protein sequences as an optimal transport problem. We have shown that with this formulation we can derive an actual distance between two sequences, as well as a "transport plan" between the two sequences that is akin to a dot plot between those sequences.

The starting point that defined the OT problem is the original problem of Monge [21]: finding a one-to-one assignment between points in a source domain and points in a target domain, knowing the cost of pairing points from the two domains. As originally phrased by Monge, however, the OT problem was deceptively simple. It proved hard to fully characterize, such as validating the existence of a solution and how this solution can be characterized. It was only when this problem was relaxed by Kantorovich [55], to the form described in Eqs. (1) and (2) (namely with a transport plan that does not require a one-to-one assignment but allows for splitting), that a better mathematical characterization was made possible. In particular, the problem could then be described as a linear problem and it became possible to prove the existence of a solution that can be characterized using techniques from convex optimization. Since Kantorovich, there have been many ways in which the OT problem has been described, sometimes simplified for specific cost functions, and analyzed (see for example [16–18,20]). It is worth mentioning for example the "invisible hand algorithm" [33], which solves the assignment problem (namely the Monge formulation of the OT problem) using a statistical physics approach similar to the one we have proposed here for the more general relaxed OT problem. Of direct relevance to our framework, however, is the entropy-regularized formulation of the OT problem proposed by Cuturi [24] that has significantly helped popularize OT and increased the range of its applications. This formulation is briefly described in Sec. II. Both the entropy-regularized OT and the statistical physics framework we have introduced considered a modified optimization problem in which the original cost function of the OT problem is either supplemented with an entropy term for the regularized OT or replaced with a physical free energy function in our formulation. The modified optimization problems are both solved over two sets of unconstrained real continuous variables, which we write here as $\lambda(k)$ and $\mu(l)$, where the indices k and l run over the points in the source domain, S_1, and target domain, S_2, respectively. While the two formulations, regularized OT and our framework, have different functionals, their solutions share a similar set of equations to describe how the continuous variables $\lambda(k)$ and $\mu(l)$ are computed. Namely, the optimal transport plan G is written as a function of the cost matrix C and of the parameters $\lambda(k)$ and $\mu(l)$,

$$G(k,l) = g(\alpha(C(k,l) + \lambda(k) + \mu(l))), \tag{43}$$

which are then computed by satisfying the constraints,

$$\sum_l G(k,l) = m_1(k), \tag{44}$$

$$\sum_k G(k,l) = m_2(l). \tag{45}$$

In both formulations, the optimal transport plan is a function of a parameter α, with $\alpha = 1/\epsilon$, the weight given to the entropic term in the regularized OT, and $\alpha = \beta = 1/(k_B T)$, i.e., the inverse of the temperature, in our statistical physics formalism. The similarities end there, and we will discuss now the differences and their impact on solving the OT problem.

First, the mapping g between the variables $\lambda(k)$ and $\mu(l)$ and the transport plan differs significantly between the two approaches, with

$$g(x) = g_1(x) = e^{-x} \tag{46}$$

for the regularized OT, and

$$g(x) = \phi(x) = \frac{e^{-x}}{e^{-x}-1} + \frac{1}{x} \tag{47}$$

for our formulation. Both mapping functions ensure that the entries $G(k,l)$ of the transport plan remain positive. The function $g_1(x)$ however is not bounded above, while the function $\phi(x)$ ensures that $G(k,l)$ belongs to [0,1]. This constraint, built in the construction of the functional free energy we have introduced, provides better control over the variations of G during the optimization. It is unclear how a similar constraint can be considered for the regularized OT problem.

Second, the finite temperature OT framework is numerically more stable than the regularized OT. The ratio of exponentials in the definition of $\phi(x)$ makes this function numerically more stable than $g_1(x)$. This question of numerical stability is of concern as the value of α is increased in an attempt to get close to the traditional OT problem (both formulations compared here map to the traditional OT problem when $\alpha \to +\infty$). For the regularized problem, log-domain stabilizations have been proposed [32], though those stabilizations still need improvement for large α, i.e., small ϵ, or small regularization. For the framework proposed here, we have run routinely computations with α (i.e., β, the inverse of temperature) on the order of 10^{11} without numerical instabilities.

The key advantage of the regularized OT formulation that it can be solved at "lightning speed," paraphrasing the title of the paper that introduced it [24]. Indeed, it can be solved with a time complexity of $O(N^2)$, compared to the $O(N^3)$ for the linear programming solution of the traditional OT problem. We have shown that the finite temperature OT problem can also be solved with a time complexity of $O(N^2)$ at each temperature. We have shown also that the procedure is similar to an annealing process as the temperature decreases, with no loss of numerical stability, or increase in computing time for very small values of the temperature. We note that there is still room for improvement. The time complexity of $O(N^2)$ of our procedure is the result of the application of an iterative conjugate gradient method for solving the linear systems that appear when resolving the SPA conditions. Our current version of this method is naive, with a simple diagonal preconditioner. We will explore more sophisticated preconditioners, as well as other iterative methods, in future work.

Formulating the OT problem with the addition of a temperature parameter has many advantages, in addition to the ones described above. In particular, it enables annealing (referred to as scaling in statistics) with respect to the temperature. While this is of advantage when solving numerically the OT problem, it also fits well with other simulation techniques such as Monte Carlo sampling to analyze for example the polytope of possible transport plan $\mathcal{G}$ and therefore recover the true values of the free energy of the system and its internal energy, as defined in Eqs. (14) and (15). We will pursue this in future studies.

Finally, we note that the OT problem considered in this paper assumes that the two sets of points considered are embedded in the same metric space, namely that we can build the cost matrix C that connect them. If those two sets of points were discrete representations of two three-dimensional shapes, it would be difficult to generate such a cost matrix between them as those shapes are not "registered"; i.e., the correspondence between the spaces in which they are embedded may not be known. Situations like this have led to an extension to the optimal transport problem with the notion of Gromov-Wasserstein distances between metric measured spaces [35]. We believe that the concept of finite temperature optimal transport can be extended in the same way into a finite temperature Gromov-Wasserstein distance. We are currently working on this problem.

ACKNOWLEDGMENTS

The work discussed here originated from a visit by P.K. at the Institut de Physique Théorique, CEA-Saclay, France. He thanks them for their hospitality and financial support. P.K. acknowledges support from NSF Award No. 1760485.

APPENDIX A: PROOF OF PROPERTY 3: METRIC PROPERTIES OF THE FREE ENERGY

We prove that the free energy defined in Eq. (14) satisfies all triangular inequalities. Let us consider three sets of points S_1, S_2, and S_3, in a metric space $\mathcal{M}$ with associated mass vectors $\mathbf{m}_1$, $\mathbf{m}_2$, and $\mathbf{m}_3$, respectively. For a pair (i,j) of those sets, we associate a cost matrix C_{ij} derived from the distance d on $\mathcal{M}$ and a transport plan polytope $\mathcal{G}(S_i,S_j)$. Recall that any matrix G_{ij} in this polytope satisfies the three conditions

$$\begin{aligned} &\forall (k,l), \quad 0 \leqslant G_{ij}(k,l) \leqslant 1, \\ &\forall k, \quad \sum_l G_{ij}(k,l) = m_i(k), \\ &\forall k, \quad \sum_k G_{ij}(k,l) = m_j(l). \end{aligned} \tag{A1}$$

The partition function for all possible transport plans between S_i and S_j is given by

$$Z_\beta(S_i,S_j) = \int_{G_{ij}\in\mathcal{G}(S_i,S_j)} d\mu_{ij} \exp[-\beta U_{ij}(G_{ij})], \tag{A2}$$

where $U_{ij}(G_{ij}) = \sum_{kl} C_{ij}(k,l) G_{ij}(k,l)$. The corresponding free energy is given by

$$\mathcal{F}_\beta(S_i,S_j) = -\frac{1}{\beta} \ln[Z_\beta(S_i,S_j)]. \tag{A3}$$

We first note that the volume of the transport plan polytope $\mathcal{G}(S_i,S_j)$ for any $(i,j) \in [1,3]^2$ is smaller than 1. Indeed, taking into account the nature of this polytope, we have

$$\begin{aligned} \int_{G\in\mathcal{G}(S_i,S_j)} d\mu_{ij} = \int_0^1 \prod_{kl} dG(k,l) \prod_k \delta\left(\sum_l G(k,l) - m_i(k)\right) \\ \times \prod_l \delta\left(\sum_k G(k,l) - m_j(l)\right). \end{aligned} \tag{A4}$$

As the $G(k,l)$ are integrated between 0 and 1, and as the constraints set by the delta functions restrain the space of possible transport plans, we have indeed that $0 \leqslant \int_{G\in\mathcal{G}(S_i,S_j)} d\mu_{ij} \leqslant 1$.

We can prove the triangular inequality of the free energy defined in Eq. (A3) using the same proof strategy as used for the standard optimal transport distance. We consider a "glued" partition function $Z^g_{beta}(S_1,S_3)$ between S_1 and S_3:

$$Z^g_\beta(S_1,S_3) = \int_{G_{12}\in\mathcal{G}(S_1,S_2)} \int_{G_{23}\in\mathcal{G}(S_2,S_3)} d\mu_{ij} d\mu_{ij} \times \exp\left(-\beta \sum_{ijk} C_{13}(i,k) \frac{G_{12}(i,j)G_{23}(j,k)}{m_2(j)}\right). \tag{A5}$$

Let $A = \sum_{ijk} C_{13}(i,k)\frac{G_{12}(i,j)G_{23}(j,k)}{m_2(j)}$. As the cost matrices are derived from the distance on the metric space in which S_1, S_2, and S_3 are embedded, we have

$$C_{13}(i,k) \leqslant C_{12}(i,j) + C_{23}(j,k), \tag{A6}$$

for all (i,j,k), and therefore,

$$A \leqslant \sum_{ijk} C_{12}(i,j)\frac{G_{12}(i,j)G_{23}(j,k)}{m_2(j)} + \sum_{ijk} C_{23}(j,k)\frac{G_{12}(i,j)G_{23}(j,k)}{m_2(j)}. \tag{A7}$$

Note that

$$\begin{aligned}\sum_{ijk} C_{12}(i,j)\frac{G_{12}(i,j)G_{23}(j,k)}{m_2(j)} &= \sum_{ij} \frac{C_{12}(i,j)G_{12}(i,j)}{m_2(j)} \sum_k G_{23}(j,k) \\ &= \sum_{ij} \frac{C_{12}(i,j)G_{12}(i,j)}{m_2(j)} m_2(j) \\ &= \sum_{ij} C_{12}(i,j)G_{12}(i,j) = U_{12}(G_{12}).\end{aligned} \tag{A8}$$

Similarly,

$$\sum_{ijk} C_{23}(j,k)\frac{G_{12}(i,j)G_{23}(j,k)}{m_2(j)} = U_{23}(G_{23}). \tag{A9}$$

Combining Eqs. (A7), (A8), and (A9), we get

$$A \leqslant U_{12}(G_{12}) + U_{23}(G_{23}), \tag{A10}$$

from which we derive

$$\exp(-\beta A) \geqslant \exp[-\beta U_{12}(G_{12})]\exp[-\beta U_{23}(G_{23})]. \tag{A11}$$

Therefore,

$$Z^g_\beta(S_1,S_3) \geqslant Z_\beta(S_1,S_2)Z_\beta(S_2,S_3). \tag{A12}$$

Note that the "glued" partition function $Z^g_\beta(S_1,S_3)$ is computed over all transport plans between S_1 and S_3 that are glued from transport plans between S_1 and S_2 and between S_2 and S_3. Those transport plans form a subset of all transport plans between S_1 and S_3. Therefore,

$$Z^g_\beta(S_1,S_3) \leqslant Z_\beta(S_1,S_3). \tag{A13}$$

Combining Eqs. (A12) and (A13), we get

$$Z_\beta(S_1,S_3) \geqslant Z_\beta(S_1,S_2)Z_\beta(S_2,S_3). \tag{A14}$$

This inequality on the partition functions translates to the following inequality for the free energy,

$$\mathcal{F}_\beta(S_1,S_3) \leqslant \mathcal{F}_\beta(S_1,S_2) + \mathcal{F}_\beta(S_2,S_3), \tag{A15}$$

which concludes the proof that $\mathcal{F}$ satisfies all triangular inequalities.

APPENDIX B: PROOF OF PROPERTY 4: MONOTONICITY OF THE FREE ENERGY AND AVERAGE ENERGY

Let us consider two sets of points S_1 and S_2 in a metric space $\mathcal{M}$ with associated mass vectors $\mathbf{m}_1$ and $\mathbf{m}_2$, respectively. We associate with this system a cost matrix C and a transport plan polytope $\mathcal{G}(S_1,S_2)$. Recall that any matrix G in this polytope satisfies the three conditions in Eq. (5). The free energy $\mathcal{F}_\beta$ of this system is related to its internal energy E_β and entropy S_β through the general relation $\mathcal{F}_\beta = E_\beta - TS_\beta$, where T is the temperature and $\beta = 1/(k_B T)$.

The internal energy is the thermodynamic average of the energy U and is given by

$$E_\beta = \langle U(G)\rangle_{G\in\mathcal{G}(S_1,S_2)} = \frac{d(\beta\mathcal{F}_\beta)}{d\beta}, \tag{B1}$$

while the entropy is given by

$$S_\beta = \beta^2 \frac{d\mathcal{F}_\beta}{d\beta} = -\frac{d\mathcal{F}_\beta}{dT}. \tag{B2}$$

An important implication of these relations is that

$$\frac{dE_\beta}{d\beta} = -(\langle U^2\rangle - \langle U\rangle^2), \tag{B3}$$

where the thermodynamics averages $\langle\ \rangle$ are computed over the polytope $\mathcal{G}(S_1,S_2)$. The quantity on the left is minus the variance of the energy and is therefore negative, for all values of β. As a result, the internal (or average) energy of the system decreases as β increases. As $U(G)$ is positive (as both the cost matrix C and the transportation plan G are positive), E_β is positive: it has a limit when $\beta \to \infty$. This limit is the traditional optimal transport distance $d(S_1,S_2)$ (see Sec. II).

The entropy is negative. Indeed, the total number of states at an energy U is given by

$$\mathcal{N}(U) = e^{S_\beta(U)} = \int_{G\in\mathcal{G}(S_i,S_j)} \delta\left(U - \sum_{kl} G(k,l)C(k,l)\right) d\mu_{ij}. \tag{B4}$$

The volume of the polytope $\mathcal{G}(S_i,S_j)$ is smaller than 1 (see Appendix A),

$$\mathcal{N}(U) = e^{S_\beta(U)} \leqslant 1, \tag{B5}$$

and therefore $S_\beta(U) \leqslant 0$. As this is true for all values of U, we have $S_\beta(T) \leqslant 0,\ \forall T$. The free energy is related to the

entropy by

$$\frac{d\mathcal{F}_\beta}{dT} = -\beta^2 \frac{d\mathcal{F}_\beta}{d\beta} = -S_\beta(T). \tag{B6}$$

Therefore

$$\frac{d\mathcal{F}_\beta}{d\beta} = \frac{S_\beta(T)}{\beta^2} \leqslant 0. \tag{B7}$$

Therefore the free energy of the system decreases as β increases. Its limit for $\beta \to \infty$ is the same as the limit of E_β, namely, the optimal transport distance $d(S_1, S_2)$.

APPENDIX C: PROOF OF THEOREM 1: CONCAVITY OF THE EFFECTIVE FREE ENERGY

We first prove that the effective free energy $F_\beta(\boldsymbol{\lambda}, \boldsymbol{\mu})$ is weakly concave, by showing that its Hessian H is negative semidefinite. H is a symmetric matrix of size $2N \times 2N$, such that its rows and columns correspond to all N λ values first, followed by all N μ values. Let ϕ' be the derivative of the function ϕ, i.e.,

$$\phi'(x) = \frac{e^{-x}}{(e^{-x}-1)^2} - \frac{1}{x^2}. \tag{C1}$$

We note first that $\phi'(x) \in [\frac{-1}{12}, 0)$, $\forall x \in \mathbb{R}$, i.e., that $\phi'(x)$ is always strictly negative. We define the matrix G' such that

$$G'(k,l) = \phi'(\beta(C(k,l) + \lambda(k) + \mu(l))). \tag{C2}$$

From Eqs. (27), we obtain

$$H(k,i) = \frac{\partial^2 F_\beta(\boldsymbol{\lambda}, \boldsymbol{\mu})}{\partial\lambda(k)\partial\lambda(i)} = \beta\delta_{ki} \sum_l G'(k,l), \tag{C3}$$

$$H(k,l) = \frac{\partial^2 F_\beta(\boldsymbol{\lambda}, \boldsymbol{\mu})}{\partial\lambda(k)\partial\mu(l)} = \beta G'(k,l), \tag{C4}$$

$$H(l,m) = \frac{\partial^2 F_\beta(\boldsymbol{\lambda}, \boldsymbol{\mu})}{\partial\mu(l)\partial\mu(m)} = \beta\delta_{lm} \sum_k G'(k,l), \tag{C5}$$

where δ are Kronecker functions, the indices k and i belong to $[1, N]$, and the indices l and m belong to $[1, N]$.

Let $\mathbf{x} = (\mathbf{x_1}, \mathbf{x_2})$ be an arbitrary vector of size $2N$. The quadratic form $Q(\mathbf{x}) = \mathbf{x}^T H \mathbf{x}$ is equal to

$$\begin{aligned} Q(\mathbf{x}) = & \sum_{i,k} x_1(k) H(k,i) x_1(i) + 2\sum_{k,l} x_1(k) H(k,l) x_2(l) \\ & + \sum_{l,m} x_2(l) H(l,m) x_2(m) \\ = & \ \beta \sum_{k,l} x_1(k)^2 G'(k,l) + 2\beta \sum_{k,l} x_1(k) G'(k,l) x_2(l) \\ & + \beta \sum_{k,l} x_2(l)^2 G'(k,l) \\ = & \ \beta \sum_{k,l} [x_1(k) + x_2(l)]^2 G'(k,l). \end{aligned} \tag{C6}$$

As $G'(k,l)$ is based on the function ϕ' that is strictly negative, the summands in the equation above are negative for all k and l, and therefore $Q(\mathbf{x})$ is negative for all vectors $\mathbf{x}$. The Hessian H is negative semidefinite. As a consequence $F_\beta(\boldsymbol{\lambda}, \boldsymbol{\mu})$ is (weakly) concave.

As $Q(\mathbf{x})$ is a sum of negative terms, it is 0 if and only if all the terms are equal to 0. This means that $\forall(k,l)$, $x_1(k) + x_2(l) = 0$. This is realized when all the coordinates to x_1 are equal and set to a parameter K, and all the coordinates to x_2 are equal and set to $-K$. Therefore 0 is an eigenvalue of H, with eigenvector $\mathbf{x} = (1, \ldots, 1, -1, \ldots, -1)$. This eigenvector corresponds to the translation invariance for the free energy. It can be removed by setting one of the parameters $\lambda(k)$ or $\mu(l)$ to zero; the free energy functional $F_\beta(\boldsymbol{\lambda}, \boldsymbol{\mu})$ on this restricted parameter space is then strictly concave.

APPENDIX D: PROOF OF PROPOSITION 5: METRIC PROPERTIES OF THE FINITE TEMPERATURE TRANSPORT ENERGY

Similarly to Appendix A, let us consider three sets of points S_1, S_2, and S_3 in a metric space $\mathcal{M}$ with associated mass vectors $\mathbf{m}_1$, $\mathbf{m}_2$, and $\mathbf{m}_3$, respectively. For a pair (i, j) of those sets, we associate a cost matrix C_{ij} derived from the distance d on $\mathcal{M}$ and a transport plan polytope $\mathcal{G}(S_i, S_j)$. Let G_{ij}^{opt} be the optimal transport plan between S_i and S_j that satisfies the saddle point equations (27) at a set value for β, and let U_{ij}^{MF} be the associated mean field optimal transport cost, i.e., $U_{ij}^{MF} = \sum_{kl} C_{ij}(k,l) G_{ij}^{\text{opt}}(k,l)$. G_{ij}^{opt} and U_{ij}^{MF} depend on β. We omit it here for clarity of presentation.

Based on the definitions above, the energy associated with G_{13}^{opt} is

$$U_{13}^{MF} = \sum_{kl} C_{13}(k,l) G_{13}^{\text{opt}}(k,l). \tag{D1}$$

As the cost matrix is metric,

$$C_{13}(k,l) \leqslant C_{12}(k,j) + C_{23}(j,l) \tag{D2}$$

for any j in $[1, N]$ (i.e., j is an index for a point in S_2), and therefore

$$\begin{aligned} U_{13}^{MF} & \leqslant \sum_{kl} C_{12}(k,j) G_{13}^{\text{opt}}(k,l) + \sum_{kl} C_{23}(j,l) G_{13}^{\text{opt}}(k,l) \\ & \leqslant \sum_k C_{12}(k,j) \sum_l G_{13}^{\text{opt}}(k,l) \\ & \quad + \sum_l C_{23}(j,l) \sum_k G_{13}^{\text{opt}}(k,l). \end{aligned} \tag{D3}$$

Note that $G_{13}^{\text{opt}}(k,l) \in \mathcal{G}(S_1, S_3)$, i.e., satisfies the transport constraints. Therefore,

$$U_{13}^{MF} \leqslant \sum_k C_{12}(k,j) m_1(k) + \sum_l C_{23}(j,l) m_3(l) \tag{D4}$$

for all $j \in [1, N]$. We rewrite this equation as

$$U_{13}^{MF} \leqslant A(j) \tag{D5}$$

for all $j \in [1, N]$, where we have defined $A(j) = \sum_k C_{12}(k,j) m_1(k) + \sum_l C_{23}(j,l) m_3(l)$.

Let us now consider the "glued" transport plan G_{13}^g defined as

$$G_{13}^g(k,l) = \sum_i \frac{G_{12}^{\text{opt}}(k,i) G_{23}^{\text{opt}}(i,l)}{m_2(i)}. \tag{D6}$$

Based on the gluing lemma, $G^g_{13} \in \mathcal{G}(S_1, S_3)$. As such,

$$m_1(k) = \sum_l G^g_{13}(k,l), \quad \forall k,$$
$$m_3(l) = \sum_k G^g_{13}(k,l), \quad \forall l. \tag{D7}$$

Replacing in the expression of $A(j)$, we get

$$\begin{aligned} A(j) &= \sum_k C_{12}(k,j) \sum_l G^g_{13}(k,l) \\ &\quad + \sum_l C_{23}(j,l) \sum_k G^g_{13}(k,l) \\ &= \sum_{k,l} [C_{12}(k,j) + C_{23}(j,l)] G^g_{13}(k,l) \\ &= \sum_{k,l} [C_{12}(k,j) + C_{23}(j,l)] \sum_i \frac{G^{\text{opt}}_{12}(k,i) G^{\text{opt}}_{23}(i,l)}{m_2(i)}. \end{aligned} \tag{D8}$$

The set of real numbers $\{C_{12}(k,j) + C_{23}(j,l)\}$ with $j \in [1,N]$ is finite. According to the well ordering principle, it has a minimum element with index j_0 such that

$$C_{12}(k,j_0) + C_{23}(j_0,l) \leqslant C_{12}(k,i) + C_{23}(i,l), \ \forall i. \tag{D9}$$

Then,

$$\begin{aligned} A(j_0) &\leqslant \sum_{k,i,l} [C_{12}(k,i) + C_{23}(i,l)] \frac{G^{\text{opt}}_{12}(k,i) G^{\text{opt}}_{23}(i,l)}{m_2(i)} \\ &\leqslant \sum_{k,i,l} C_{12}(k,i) \frac{G^{\text{opt}}_{12}(k,i) G^{\text{opt}}_{23}(i,l)}{m_2(i)} \\ &\quad + \sum_{k,i,l} C_{23}(i,l) \frac{G^{\text{opt}}_{12}(k,i) G^{\text{opt}}_{23}(i,l)}{m_2(i)} \leqslant U^{MF}_{12} + U^{MF}_{23}. \end{aligned} \tag{D10}$$

Since $U^{\text{opt}}_{13} \leqslant A(j)$ for all j [Eqn. (D5)], $U^{\text{opt}}_{13} \leqslant A(j_0)$, and therefore

$$U^{MF}_{13} \leqslant U^{MF}_{12} + U^{MF}_{23}, \tag{D11}$$

which concludes the proof that U^{MF} satisfies all triangular inequalities.

APPENDIX E: PROOF OF PROPOSITION 6: MONOTONICITY AND LIMITS OF $F^{MF}(\beta)$ AND $U^{MF}(\beta)$

Let us consider two sets of points S_1 and S_2 in a metric space $\mathcal{M}$ with associated mass vectors $\mathbf{m}_1$ and $\mathbf{m}_2$, respectively. We associate with this system a cost matrix C and a transport plan polytope $\mathcal{G}(S_1, S_2)$. In Appendix B we have established that the exact free energy and internal energy defined in Eqs. (14) and (15), respectively, are monotonic functions of the parameter β, and converge to the actual optimal transport distance $d(S_1, S_2)$ when $\beta \to \infty$. Here we consider the approximation of those quantities obtained with the saddle point approximation, namely the mean field values F^{MF} and U^{MF}, and show that they satisfy the same properties.

The effective free energy $F_\beta(\boldsymbol{\lambda}, \boldsymbol{\mu})$ defined in Eq. (22) is a function of the cost matrix C and of real unconstrained variables $\lambda(k)$ and $\mu(l)$. For sake of simplicity, for any $(k,l) \in [1,N]^2$, we define

$$x_{kl} = C(k,l) + \lambda(k) + \mu(l). \tag{E1}$$

The effective free energy is then

$$\begin{aligned} F_\beta(\boldsymbol{\lambda}, \boldsymbol{\mu}) &= -\left[\sum_k \lambda(k) m_1(k) + \sum_l \mu_l m_2(l) \right] \\ &\quad - \frac{1}{\beta} \sum_{kl} \ln \left(\frac{1 - e^{-\beta x_{kl}}}{\beta x_{kl}} \right). \end{aligned} \tag{E2}$$

As written above, $F_\beta(\boldsymbol{\lambda}, \boldsymbol{\mu})$ is a function of the independent variables β, $\lambda(k)$, and $\mu(l)$. However, under the saddle point approximation, the variables $\lambda(k)$ and $\mu(l)$ are constrained by the conditions

$$\begin{aligned} \frac{\partial F_\beta(\boldsymbol{\lambda}, \boldsymbol{\mu})}{\partial \lambda(k)} &= 0, \\ \frac{\partial F_\beta(\boldsymbol{\lambda}, \boldsymbol{\mu})}{\partial \mu(l)} &= 0, \end{aligned} \tag{E3}$$

and the free energy under those constraints is written as $F^{MF}(\beta)$. In the following, we will use the notations $\frac{dF^{MF}(\beta)}{d\beta}$ and $\frac{\partial F^{MF}(\beta)}{\partial \beta}$ to differentiate between the total derivative and partial derivative of $F^{MF}(\beta)$ with respect to β, respectively. Based on the chain rule,

$$\begin{aligned} \frac{dF^{MF}(\beta)}{d\beta} &= \frac{\partial F_\beta(\boldsymbol{\lambda}, \boldsymbol{\mu})}{\partial \beta} + \sum_k \frac{\partial F_\beta(\boldsymbol{\lambda}, \boldsymbol{\mu})}{\partial \lambda(k)} \frac{\partial \lambda(k)}{\partial \beta} \\ &\quad + \sum_l \frac{\partial F_\beta(\boldsymbol{\lambda}, \boldsymbol{\mu})}{\partial \mu(l)} \frac{\partial \mu(l)}{\partial \beta}. \end{aligned} \tag{E4}$$

Using the constraints (E3), we find that

$$\frac{dF^{MF}(\beta)}{d\beta} = \frac{\partial F_\beta(\boldsymbol{\lambda}, \boldsymbol{\mu})}{\partial \beta}, \tag{E5}$$

namely that the total derivative with respect to β is in this specific case equal to the corresponding partial derivative, which is easily computed to be

$$\frac{dF^{MF}(\beta)}{d\beta} = \frac{1}{\beta^2} \sum_{kl} \left[\ln \left(\frac{1 - e^{-\beta x_{kl}}}{\beta x_{kl}} \right) + \beta x_{kl} \phi(\beta x_{kl}) \right], \tag{E6}$$

where $\phi(x) = \frac{e^{-x}}{e^{-x}-1} + \frac{1}{x}$, as defined in Eq. (28). Let $f(x) = \ln(\frac{1-e^{-x}}{x}) + x\phi(x)$. In Fig. 5, we represent the two functions $\phi(x)$ and $f(x)$. As mentioned in the main text of the paper, $\phi(x)$ is monotonically constrained in the interval [0,1] and therefore correctly represents the possible values for the corresponding transport plan. The function $f(x)$ is continuous and defined over all real values x [with the extension $f(0) = 0$] and is bounded above by 0, i.e., $f(x) \leqslant 0, \ \forall x \in \mathbb{R}$.

As

$$\frac{dF^{MF}(\beta)}{d\beta} = \frac{1}{\beta^2} \sum_{kl} f(\beta x_{kl}), \tag{E7}$$

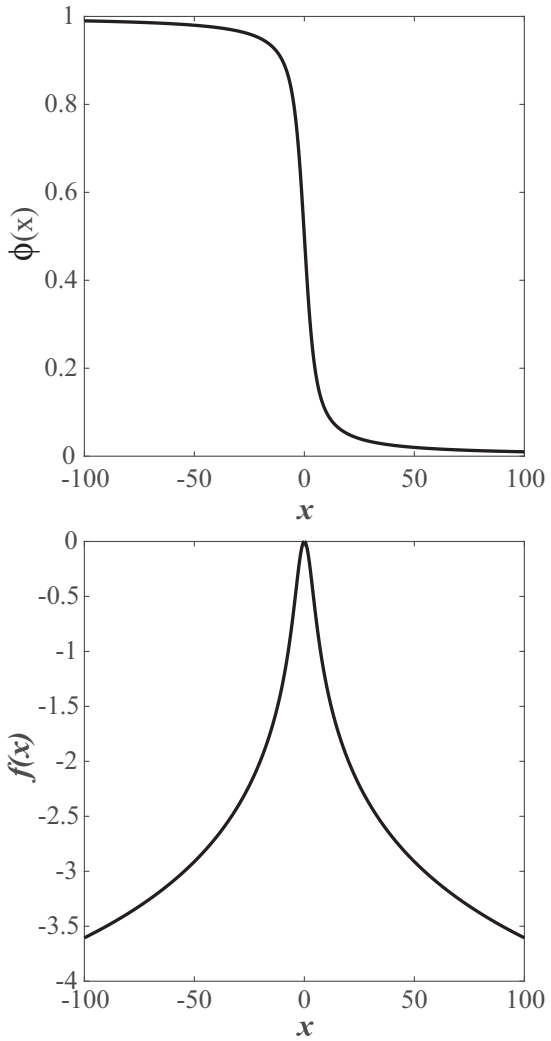

FIG. 5. The two functions $\phi(x)$ and $f(x)$.

we conclude that

$$\frac{dF^{MF}(\beta)}{d\beta} \leqslant 0, \tag{E8}$$

namely that $F^{MF}(\beta)$ is a monotonically decreasing function of β. In addition, we note that $F^{MF}(\beta)$ is the mean field approximation of the true free energy $\mathcal{F}_\beta$ and that this approximation becomes exact when β tends to ∞. Therefore,

$$\lim_{\beta\to\infty} F^{MF}(\beta) = \lim_{\beta\to\infty} \mathcal{F}(\beta) = d(S_1, S_2), \tag{E9}$$

where $d(S_1, S_2)$ is the traditional optimal transport distance between the two sets of points S_1 and S_2.

Let

$$U_\beta(\boldsymbol{\lambda}, \boldsymbol{\mu}) = \sum_{kl} C_{kl} G_\beta(k, l) \tag{E10}$$

and the corresponding mean field approximation of the internal energy at the saddle point

$$U^{MF}(\beta) = \sum_{kl} C_{kl} G_\beta^{\text{opt}}(k, l). \tag{E11}$$

Before computing $\frac{dU^{MF}(\beta)}{d\beta}$, let us first notice that by replacing Eq. (E2) into (E6), we get

$$\beta \frac{dF^{MF}(\beta)}{d\beta} = -F^{MF}(\beta) - \sum_k \lambda(k) m_1(k) - \sum_l \mu_l m_2(l) + \sum_{kl} x_{kl} \phi(\beta x_{kl}). \tag{E12}$$

F^{MF} is the value of the free energy at the saddle point of the free energy functional and is associated with a transport plan G_β^{opt} that satisfies Eqs. (27). Therefore,

$$\beta \frac{dF^{MF}(\beta)}{d\beta} = -F^{MF}(\beta) - \sum_{kl} \lambda(k) G_\beta^{\text{opt}}(k, l) - \sum_{kl} \mu_l G_\beta^{\text{opt}}(k, l) + \sum_{kl} x_{kl} G_\beta^{\text{opt}}(k, l). \tag{E13}$$

Therefore,

$$\begin{aligned} \beta \frac{dF^{MF}(\beta)}{d\beta} &= -F^{MF}(\beta) - \sum_{kl} [x_{kl} - \lambda(k) - \mu(l)] G_\beta^{\text{opt}}(k, l) \\ &= -F^{MF}(\beta) + \sum_{kl} C(k, l) G_\beta^{\text{opt}}(k, l) \\ &= -F^{MF}(\beta) + U^{MF}(\beta). \end{aligned} \tag{E14}$$

Note that this equation can be rewritten as

$$\begin{aligned} U^{MF}(\beta) &= F^{MF}(\beta) + \beta \frac{dF^{MF}(\beta)}{d\beta} \\ &= \frac{d(\beta F^{MF}(\beta))}{d\beta}; \end{aligned} \tag{E15}$$

i.e., it extends the relationship (B1) known between the true free energy and the average energy to their mean field counterparts.

Based on the chain rule,

$$\frac{dU^{MF}(\beta)}{d\beta} = \frac{\partial U_\beta(\boldsymbol{\lambda}, \boldsymbol{\mu})}{\partial \beta} + \sum_k \frac{\partial U_\beta(\boldsymbol{\lambda}, \boldsymbol{\mu})}{\partial \lambda(k)} \frac{\partial \lambda(k)}{\partial \beta} + \sum_l \frac{\partial U_\beta(\boldsymbol{\lambda}, \boldsymbol{\mu})}{\partial \mu(l)} \frac{\partial \mu(l)}{\partial \beta}. \tag{E16}$$

Let us compute all partial derivatives in this equation:

$$\begin{aligned} \frac{\partial U_\beta(\boldsymbol{\lambda}, \boldsymbol{\mu})}{\partial \lambda(k)} &= \frac{\partial F_\beta(\boldsymbol{\lambda}, \boldsymbol{\mu})}{\partial \lambda(k)} + \beta \frac{\partial}{\partial \lambda(k)} \left(\frac{\partial F_\beta(\boldsymbol{\lambda}, \boldsymbol{\mu})}{\partial \beta} \right) \\ &= \frac{\partial F_\beta(\boldsymbol{\lambda}, \boldsymbol{\mu})}{\partial \lambda(k)} + \beta \frac{\partial}{\partial \beta} \left(\frac{\partial F_\beta(\boldsymbol{\lambda}, \boldsymbol{\mu})}{\partial \lambda(k)} \right) = 0, \end{aligned} \tag{E17}$$

where the zero is a consequence of the SPA constraints. Similarly,

$$\begin{aligned} \frac{\partial U_\beta(\boldsymbol{\lambda}, \boldsymbol{\mu})}{\partial \mu(l)} &= \frac{\partial F_\beta(\boldsymbol{\lambda}, \boldsymbol{\mu})}{\partial \mu(l)} + \beta \frac{\partial}{\partial \mu(l)} \left(\frac{\partial F_\beta(\boldsymbol{\lambda}, \boldsymbol{\mu})}{\partial \beta} \right) \\ &= \frac{\partial F_\beta(\boldsymbol{\lambda}, \boldsymbol{\mu})}{\partial \mu(l)} + \beta \frac{\partial}{\partial \beta} \left(\frac{\partial F_\beta(\boldsymbol{\lambda}, \boldsymbol{\mu})}{\partial \mu(l)} \right) = 0. \end{aligned} \tag{E18}$$

Finally,

$$\begin{aligned}\frac{\partial U_\beta(\boldsymbol{\lambda},\boldsymbol{\mu})}{\partial \mu(l)} &= 2\frac{\partial F_\beta(\boldsymbol{\lambda},\boldsymbol{\mu})}{\partial \beta} + \beta\frac{\partial}{\partial \beta}\left(\frac{\partial F_\beta(\boldsymbol{\lambda},\boldsymbol{\mu})}{\partial \beta}\right)\\ &= 2\frac{\partial F_\beta(\boldsymbol{\lambda},\boldsymbol{\mu})}{\partial \beta}\\ &\quad + \beta\left(\frac{-2}{\beta}\frac{\partial F_\beta(\boldsymbol{\lambda},\boldsymbol{\mu})}{\partial \beta} + \frac{1}{\beta^2}\sum_{kl}\beta x_{kl}^2\phi'(\beta x_{kl})\right)\\ &= \sum_{kl}\beta x_{kl}^2\phi'(\beta x_{kl}). \end{aligned} \tag{E19}$$

As x_{kl}^2 is always positive and $\phi'(x)$ is always negative, we have

$$\frac{dU^{MF}(\beta)}{d\beta} = \frac{\partial U_\beta(\boldsymbol{\lambda},\boldsymbol{\mu})}{\partial \mu(l)} \leqslant 0, \tag{E20}$$

and the function $U^{MF}(\beta)$ is a monotonically decreasing function of β. In addition, we note that $U^{MF}(\beta)$ is the mean field approximation of the true internal energy E_β and that this approximation becomes exact when β tends to ∞. Therefore,

$$\lim_{\beta\to\infty} U^{MF}(\beta) = \lim_{\beta\to\infty} E(\beta) = d(S_1, S_2), \tag{E21}$$

where $d(S_1, S_2)$ is the traditional optimal transport distance between the two sets of points S_1 and S_2.

[1] S. Kolouri, S. R. Park, M. Thorpe, D. Slepcev, and G. K. Rohde, IEEE Signal Process. Mag. **34**, 43 (2017).

[2] Y. Rubner, C. Tomasi, and L. Guibas, in *Proceedings of the Sixth International Conference on Computer Vision* (IEEE, 1998), pp. 59–66.

[3] Y. Rubner, C. Tomasi, and L. Guibas, Int. J. Comput. Vision **40**, 99 (2000).

[4] S. Haker, L. Zhu, A. Tannenbaum, and S. Angenent, Int. J. Comput. Vision **60**, 225 (2004).

[5] L.-P. Saumier, M. Agueh, and B. Khouider, arXiv:1009.6039.

[6] S. Kolouri, A. Tosun, J. Ozolek, and G. Rohde, Pattern Recognit. **51**, 453 (2016).

[7] G. Huang, C. Guo, M. Kusner, Y. Sun, F. Sha, and K. Weinberger, in *Advances in Neural Information Processing Systems 30* (NIPS Foundation, 2016), pp. 4862–4870.

[8] A. Rolet, M. Cuturi, and G. Peyré, in *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics* (AISTATS, 2016), pp. 630–638.

[9] N. Guillen and R. McCaan, arXiv:1011.2911.

[10] A. Figalli and C. Villani, in *Nonlinear PDE's and Applications* (Springer, Berlin, 2011), pp. 171–217.

[11] Z. Su, Y. Wang, R. Shi, W. Zeng, J. Sun, F. Luo, and X. Gu, IEEE Trans. Pattern Anal. Mach. Intell. **37**, 2246 (2015).

[12] M. Ma, N. Lei, W. Chen, K. Su, and X. Gu, Graph. Models **90**, 13 (2017).

[13] R. Flamary, C. Févotte, N. Courty, and V. Emyia, in *Advances in Neural Information Processing Systems 30* (NIPS Foundation, 2016), pp. 703–711.

[14] G. Schiebinger, J. Shu, M. Tabaka, B. Cleary, V. Subramanian, A. Solomon, J. Gould, S. Liu, S. Lin, P. Berube, L. Lee, J. Chen, J. Brumbaugh, P. Rigollet, K. Hochedlinger, R. Jaenisch, A. Regev, and E. Lander, Cell **176**, 928 (2019).

[15] P. Dixit and K. Dill, J. Chem. Phys. **150**, 054105 (2019).

[16] C. Villani, *Topics in Optimal Transport* (American Mathematical Society, Providence, RI, 2003).

[17] C. Villani, *Optimal Transport: Old and New*, Grundlehren der mathematischen Wissenschaften, Vol. 338 (Springer-Verlag, Berlin, Heidelberg, 2008), 1st ed.

[18] F. Santambrogio, *Optimal Transport for Applied Mathematicians*, Progress in Nonlinear Differential Equations and Their Applications, Vol. 87 (Birkhäuser, Basel, 2015), 1st ed.

[19] B. Levy and E. Schwindt, Comput. Graph. **72**, 135 (2018).

[20] G. Peyré and M. Cuturi, Found. Trends Mach. Learn. **11**, 355 (2019).

[21] G. Monge, Histoire de l'Académie Royale des Sciences et Mémoires de Mathématique et de Physique tirés des registres de cette Académie **1781**, 666 (1781).

[22] C. Léonard, Discrete Continuous Dyn. Syst. Series A **34**, 1533 (2014).

[23] G. Sierksma and Y. Zwols, *Linear and Integer Optimization: Theory and Practice*, 3rd ed. (CRC Press, New York, 2015).

[24] M. Cuturi, in *Advances in Neural Information Processing Systems 26*, edited by C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger (Curran Associates, Inc., 2013), pp. 2292–2300.

[25] W. E. Deming and F. F. Stephan, Ann. Math. Stat. **11**, 427 (1940).

[26] R. Sinkhorn, Ann. Math. Stat. **35**, 876 (1964).

[27] R. Sinkhorn and P. Knopp, Pacific J. Math. **21**, 343 (1967).

[28] J. Benamou, G. Carlier, M. Cuturi, L. Nenna, and G. Peyré, SIAM J. Sci. Comput. **37**, A1111 (2015).

[29] A. Genevay, M. Cuturi, G. Peyré, and F. Bach, in *Advances in Neural Information Processing Systems 29* (Curran Associates, Inc., 2016), pp. 3440–3448.

[30] P. Dvurechensky, A. Gasnikov, and A. Kroshnin, in *35th International Conference on Machine Learning (ICML 2018)* (Curran Associates, Inc., 2018), pp. 1367–1376.

[31] L. Chizat, G. Peyré, B. Schmitzer, and F.-X. Vialard, Math. Comput. **87**, 2563 (2018).

[32] B. Schmitzer, SIAM J. Sci. Comput. **41**, A1443 (2019).

[33] J. Kosowsky and A. Yuille, Neural Networks **7**, 477 (1994).

[34] P. Koehl, M. Delarue, and H. Orland, Phys. Rev. Lett. **123**, 040603 (2019).

[35] F. Mémoli, in *Eurographics Symposium on Point-Based Graphics* (Eurographics, 2007), pp. 81–90.

[36] M. Waterman, *Introduction to Computational Biology: Maps, Sequences, and Genomes* (Chapman and Hall/CRC Interdisciplinary Statistics, Boca Raton, FL, 1995).

[37] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison, *Biological Sequence Analysis: Probabilistic Models of Nucleic Acids and Proteins* (Cambridge University Press, New York, 1998).

[38] D. Gusfield, *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology* (Cambridge University Press, New York, 1997).

[39] S. Vinga and J. Almeida, Bioinformatics **19**, 513 (2003).
[40] O. Bonham-Carter, J. Steele, and D. Bastola, Brief Bioinform. **15**, 890 (2014).
[41] S. Vinga, Brief Bioinform. **15**, 341 (2014).
[42] A. Zielezinski, S. Vinga, J. Almeida, and W. Karlowski, Genome Biol. **18**, 186 (2017).
[43] I. Schwende and T. Pham, Brief Bioinform. **15**, 354 (2014).
[44] S. Henikoff and J. Henikoff, Proc. Natl. Acad. Sci. USA **89**, 10915 (1992).
[45] W.-J. Shen, H.-S. Wong, Q.-W. Xiao, X. Guo, and S. Smale, Found. Comput. Math. **14**, 951 (2013).
[46] N. Fox, S. Brenner, and J.-M. Chandonia, Nucl. Acids. Res. **42**, D304 (2014).
[47] W. Pearson and D. Lipman, Proc. Natl. Acad. Sci. USA **85**, 2444 (1988).
[48] J. Bray and J. Curtis, Ecol. Monogr. **27**, 325 (1957).
[49] I. Ladunga, Bioinformatics **15**, 1028 (1999).
[50] T. Smith and M. Waterman, J. Molec. Biol. **147**, 195 (1981).
[51] S. Needleman and C. Wunsch, J. Molec. Biol. **48**, 443 (1970).
[52] A. Gibbs and G. McIntyre, Eur. J. Biochem. **16**, 1 (1970).
[53] E. Anderson, Z. Bai, C. Bischof, J. Demmel, J. Dongarra, J. Ducroz, A. Greenbaum, S. Hammarling, A. Mckenney, and D. Sorensen, LAPACK: A Portable Linear Algebra Library for High-Performance Computers, Tech. Rep. CS-90-105, Computer Science Department, University of Tennessee, Knoxville, TN, 1990 .
[54] A. Thibault, L. Chizat, C. Dossal, and N. Papadakis, in *NIPS'17 Workshop on Optimal Transport and Machine Learning* (NIPS Foundation, 2017).
[55] L. V. Kantorovich, Dokl. Akad. Nauk SSSR **37**, 227 (1942) [J. Math. Sciences **133**, 1381 (2006)].